Author:
Demiralp, Merve

Title:
The C-test as a Second Language Proficiency Estimate and Screening Test in Turkish

An Argument-based Validation Study

The C-test as a Second Language Proficiency Estimate and Screening Test in Turkish: An Argument-based Validation Study

Merve Demiralp

School of Education
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Social Sciences and Law.

March, 2020

Word Count: 74,055 (excluding references and appendices)

ABSTRACT

Assessing the language proficiency of second language (L2) learners and bilinguals is essential in second language acquisition (SLA) research; to control for language proficiency or to select participants into an SLA study (i.e., Gaillard, 2015; Drackert, 2016; Norris & Ortega, 2012; Tremblay, 2011). L2 proficiency testing is also important in educational programs to make decisions such as placing students into appropriate levels of language programs and screening students to identify those with limited language skills (i.e., Elder & von Randow, 2008; Norris, 2006, 2008). At present, there is an insufficient body of standardized and validated measurement instruments in SLA research (Norris & Ortega, 2012). Similarly, validating low-stakes screening and diagnostic tests has been underestimated in educational context despite their impact when they are published online open to hundreds of thousands of learners (i.e., Alderson et al, 2015; Schmidgall et al, 2017).

To address these issues, this present research developed a Turkish C-test for adult L2 learners of Turkish and validated it by using Kane's argument-based approach (2006) for two different uses: (1) SLA research purposes where researchers need to control the language proficiency of their research participants (study 1); (2) educational purposes as a screening test for the Turkish Proficiency Exam (TYS) used to admit Turkish L2 learners into Turkish-medium universities (study 2). Both validation studies adopted a mixed-methods approach in order to gain better insight into stakeholders' perception of the uses of the Turkish C-test.

For study 1, the Turkish C-test was administered to 85 Turkish L2 learners in the UK and USA along with a background questionnaire and feedback survey. This was augmented with a second survey administered to 10 SLA researchers, and interviews were conducted with five of these researchers. The Turkish C-test was found to discriminate between four different levels of Turkish L2 learners with an IRT separation reliability of .94. Furthermore, the internal consistency of the texts was high with a reliability value of .92, and texts

functioned consistently across both UK and USA settings. Regarding stakeholders' perceptions of the test, although they found the test practical, they were sceptical towards what it can measure and be used for.

For study 2, the Turkish C-test was administered to 79 TYS candidates alongside a background questionnaire and feedback survey. Interviews were also conducted with 13 of these participants. This was augmented with a second survey administered to 34 instructors of Turkish, and interviews were conducted with two of these instructors. The Turkish C-test was found to moderately to strongly correlate with each TYS section (reading, writing, listening, oral) as well as TYS total score. It was also successful in placing 68% of the TYS candidates in the right TYS levels although it couldn't discriminate between C1 and C2 levels. Qualitative data suggested that test takers and instructors were sceptical about the relevance of the Turkish C-test to the spoken sections of TYS despite the strong quantitative findings. Nevertheless, test takers reported that the C-test helped them understand their need to learn more and thus would be useful in exam preparation for TYS. Overall, the findings suggest that the Turkish C-test can predict success or failure in TYS and could therefore be used as a screening test. This can help TYS candidates save time, money, and energy.

This dissertation is unique in showing the development stages of a Turkish C-test step by step with language specific factors. Through an argument-based approach to validation, it provided researchers and learners with a freely available Turkish C-test that can be effectively used for research and screening purposes on the condition that findings are replicated with a follow-up study. If future researchers or practitioners wanted to use a Turkish C-test for different populations or uses, they can follow the steps and guidance stated in this dissertation while developing their own C-tests.

ACKNOWLEDGEMENTS

There is no joy without gratitude. I am very grateful for the support of numerous people and institutions, which made the completion of this dissertation possible.

Above all, I've been very fortunate to have two remarkable scholars, Shelley McKeown Jones and George Leckie, in my supervision team. Thank you both for seeing me through the most challenging times of doing this PhD. I very much appreciate your endless support as well as the substantial time and energy you dedicated to reading many drafts of this work. I also want to thank my MA supervisor John Norris for raising my interest in language assessment and guiding me to the PhD path. It has been a real privilege to have worked with him who is not only a great researcher and team leader, but also a very inspiring teacher.

Many thanks to George Leckie, William Browne, and Liz Washbrook for the excellent statistical training they provide for graduate students in our program. I've been very fortunate to join your classes both as a student and a teaching assistant. Many thanks to my examiners Liz Washbrook and Anthony Green for a very positive viva experience and the constructive feedback they have given. I have been privileged to have had such highly respected scholars in my viva.

This dissertation would not be possible without the graduate studies funding I received from the Turkish Ministry of Education. I also acknowledge the University of Bristol School of Education and Alumni Foundation for financially supporting me to present my work in conferences. I am very thankful to the Educational Testing Service for awarding me the TOEFL small grants for doctoral researchers, which helped me in data collection. I also want to thank Yunus Emre Institute, and particularly Nursel Tan Elmas, for helping me reach participants and sharing information about the Turkish Proficiency Exam. Many thanks to the Turkish language departments of several UK and US universities for helping me in data collection. I appreciate the contributions of all my participants, including instructors, researchers, learners, and speakers of Turkish, teşekkürler!

A huge thanks to Nihal Çalışkan, Özgü Güntekin, and James Dirgin for the fruitful discussions, phone and Skype calls we had on the linguistics of the Turkish language.

Thanks to my fantastic friends for sticking with me although I often failed to stay in touch: Belma, Duygu, Jeongeoun (Janey), Pooneh and Andy, Diana, Katy, Laura, Sweeney, Pınar, İnci, Aslıhan, and Özlem. Special thanks to my sidekick Tiny for reminding me to take regular walks and be active while writing my dissertation.

I also want to thank my parents for their support, understanding, and belief in me during my studies. Last but not least, I owe the greatest gratitude to my companion star Jacek. Dziękuję Ci Jacusiu for always being there for me and taking all my PhD moods like a zen master. I wouldn't be able to complete this chapter of my life without you. Kołyszesz moim światem!

DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

The findings from the test development chapter was published as: Demiralp, M. (2018). Designing a C test for foreign language learners of Turkish. In J. M. Norris (Ed.), Developing C-tests for estimating proficiency in foreign language research. Frankfurt am Main, Germany: Peter Lang. DOI: https://doi.org/10.3726/b13235

SIGNED: DATE:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CTT – Classical Test Theory

CEFR – The Common European Framework of Reference for Languages

DIF – Differential Item Functioning

EIT – Elicited Imitation Test

ELBA – English Language Battery

ELT – English Language Teaching

EPT – English Placement Test

ESL – English as a Second Language

EQ – Evaluation Question

FL – Foreign Language

IELTS – The International English Language Testing System

ILR – Interagency Language Roundtable

IRT – Item Response Theory

LCTL – Less Commonly Taught Languages

L1 – First Language

L2 – Second Language

L3 – Third Language

MnSq – Mean Square

OnDaF – Online-einstufungstest Deutsch als Fremdsprache; Online Placement Test of German as a Foreign Language

PCM – Partial Credit Model

RSM – Rating Scale Model

SD – Standard Deviation

SLA – Second Language Acquisition

TestDaF – Test Deutsch als Fremdsprache; Test of German as a Foreign Language

TOEIC – The Test of English for International Communication

TOEFL – Test of English as a Foreign Language

TUD –Türkçe Ulusal Derlemi; Turkish National Corpus

TYS – Türkçe Yeterlilik Sınavı; Turkish Proficiency Exam

YEE – Yunus Emre Enstitüsü; Yunus Emre Institute

YÖK – Yükseköğretim Kurulu; Council of Higher Education of Turkey

CHAPTER 1: INTRODUCTION

1.1 Assessment of Proficiency in SLA Research and Educational Programs

Assessing the language proficiency of second language[1] (L2) learners and bilinguals is essential in second language acquisition (SLA) research for various reasons. These reasons include controlling for language proficiency and selecting participants into an SLA study (i.e., Gaillard, 2015; Drackert, 2016; Norris & Ortega, 2012; Tremblay, 2011). L2 proficiency testing is also important in educational programs to make decisions such as placing students into appropriate levels of language programs and screening students to identify those with limited language skills (i.e., Elder & von Randow, 2008; Green, 2012; Norris, 2006, 2008). While SLA assessment focuses on L2 knowledge constructs that researchers want to find out such as language proficiency, educational assessment depends on decisions and consequences directly affecting test takers such as placement into the correct level of language classrooms.

L2 assessment should be done through a systematic and replicable technique that enables researchers to observe, elicit and interpret learner data in SLA research (Norris & Ortega, 2012). Using such a systematic and replicable proficiency measurement ensures generalizability, replicability and interpretability across different contexts. However, at present, there is an insufficient body of standardized and validated measurement instruments in SLA research which would make the generalizability of SLA research findings easier (Norris & Ortega, 2012). The standardized tests of language proficiency such as Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS) are

[1] In this paper, second language is used as an umbrella term for languages learned both abroad as a foreign language and in the target language community as a second language.

expensive, time-consuming, and thus, oftentimes impractical to use in SLA research. This situation becomes more problematic in less commonly taught languages (LCTL) such as Turkish since most assessment related research is done in more commonly used languages although a standardised Turkish proficiency test is needed by Turkish SLA researchers (i.e., Gürel, 2016, see section 3.2.1 for details). As a result of these reasons, most SLA studies do not measure language proficiency through systematic methods, and some studies solely depend on factors such as institutional status or year of study to determine L2 learners' proficiency (Hulstijn, 2012; Tremblay, 2011). However, these indicators of L2 proficiency are not systematic or replicable and do not provide generalization across studies. Therefore, in SLA research, there is a need for "short-cut" measurements "that can within relatively few items and short test-administration time, provide reliable and accurate estimations of holistic language proficiency across a broad range of levels" (Norris, 2018, p. 11).

Regarding educational contexts, L2 proficiency assessment might be used for multiple purposes including the following: (1) making decisions about learners such as admission, enrolment, and placement tests; (2) informing teachers and learners about learners' language ability, progress, and needs such as diagnostic and screening tests; (3) evaluating language programs such as program review and institutional accreditation (Norris, 2008). Short-cut estimates of language proficiency are useful in the context of low-stakes screening tests to provide information to teachers and learners in a quick way (see section 2.2.3). For example, screening tests can help learners to decide whether their level is appropriate for a high-stakes proficiency test and direct them to the right level test. (i.e., Schmidgall, Getman, & Zu, 2017; Stansfield & Hewitt, 2005). Thus, they can be low-cost tools by preventing time and money waste on a test inappropriate for a candidate's level (Schmidgall et al, 2017).

Although there is an extensive body of research regarding the validation of high stakes assessments (i.e., TOEFL, IELTS), there is relatively little research on the evaluation of low-stakes screening tests. For example, Cambridge English Placement Test and The Exam English level tests, both of which are online and quick measurements directing learners to higher-stakes exams by providing them an estimate about their levels according to the Common European Framework of Reference for Languages (CEFR), are not clearly backed by research or formal validation (Schmidgall et al, 2017). However, validation of screening tests is also necessary given that these tests are published online, are open to thousands of learners, and provide immediate information about their proficiency levels (Chapelle, Jamieson, & Hegelheimer, 2003)

1.2 C-test as an Estimate of General Language Proficiency

As a direct consequence to the lack of validated short-cut estimates of language proficiency, there have been attempts to validate alternative measurements of language proficiency in various languages (i.e., Gaillard, 2014; Drackert, 2016; Norris, 2018; Tremblay, 2011). Reduced redundancy test types (see section 2.3 for details) are suggested as short-cut estimates of language proficiency (i.e., Norris, 2018; Tremblay, 2011), and they measure L2 learners' ability to function in an L2 under reduced redundancy conditions such as deleting some portions of a text and adding noise to speech utterances (Spolsky, 1969, 1973). Examples include the cloze test, the elicited imitation test (EIT), and the C-test. The C-test and cloze test are written reduced redundancy tests providing good estimates of general (global) language proficiency (see section 2.3.2 for details about the preference of C-tests over cloze tests), whereas the EIT is a spoken test measuring oral language proficiency (Drackert, 2016). The present research focuses on the C-test since it is a promising

area in the assessment of LCTL and can provide an estimate of general language proficiency (see Grotjahn, 2017 for an electronic version of the latest C-test bibliography[2]). In laymen's terms, the C-test involves students reading short texts where various words have been deleted and completing the gaps. C-tests have several advantages. First, they are very practical given the ease of development, administration (in paper-pencil or online format) and scoring in a short amount of time. They typically consist of four to six short texts with 20 or 25 gaps in each (Raatz & Klein-Braley, 1985) and can be completed within 20 minutes for a typical 4-text C-test (i.e., 5 minutes per text). Second, they are sufficiently global in terms of language abilities and knowledge (Tremblay, 2011). Third, they meet validity and reliability standards while discriminating between different levels of L2 learners. However, it is less clear whether C-tests can distinguish among high-level proficient learners or whether they work best at lower ranges of proficiency (see section 2.3.2.2).

There is an extensive body of research about the uses of C-tests in some European languages, and C-tests are commonly used in mainstream testing in German. For example, Online-Spracheinstufungstest[3] (onSET), which is an online German L2 proficiency test consisting of 8 C-test texts, is a screening test for Test Deutsch als Fremdsprache (TestDaF; Test of German as a Foreign Language), a standardised university language entrance test which consists of reading, writing, listening and speaking sections. In other words, onSET is used to see whether students are ready to take TestDaF (Eckes, 2014). The uses of C-tests in non-European languages, such as Turkish, should be examined since they are a promising way of estimating proficiency in a short amount of time.

2 http://www.c-test.de/
3 www.onset.de

1.3 Research Gaps and Study Purpose

To date, there is an insufficient body of validated measurement instruments, that can be quickly and effectively implemented, both in SLA research and educational contexts. C-tests have been suggested as a solution to address this issue (Norris, 2018). However, most C-test related research has been done in European languages, and there are only two studies in Turkish among 567 entries in the latest C-test bibliography (see Grotjahn, 2017). These two studies focus on bilingual pupils and do not involve L2 learners with different proficiency levels (e.g., Daller *et al*, 2002; Baur & Meder, 1994; Caprez & Gönç, 2006). Furthermore, they focus mainly on the development part of the test and do not provide any information about the validation of the test for different uses such as screening or placement. There is therefore a clear need for a standardized and accessible Turkish language proficiency test, that fits within researchers' time limitations such as a C-test, to generalize the results across studies in Turkish SLA (i.e., Gürel, 2016).

Regarding educational contexts of Turkish L2, the number of Turkish L2 learners studying at Turkish-medium universities has almost tripled within the last decade according to the 2018 statistics[4] reported by the Council of Higher Education of Turkey (YÖK[5]) (see section 3.2 for details). In order to facilitate the admission and enrolment of international students into these universities, the *Turkish Proficiency Exam* (TYS[6]) was developed (see section 3.2.2 for details about TYS). Given the costs of taking TYS and the need to pass it to study at Turkish universities, students and educators are in urgent need of a low-cost screening test such as C-test that would

4 https://istatistik.yok.gov.tr/
5 YÖK stands for Yükseköğretim Kurulu
6 TYS stands for Türkçe Yeterlilik Sınavı

help to determine whether a student is ready to take TYS, which was also supported by interviews with students (see section 7.7.5.2).

Addressing an important gap in the literature, the present research aims to develop a Turkish C-test for adult L2 learners of Turkish and validate its uses for both SLA research purposes where researchers need to control the language proficiency of their research participants and for educational purposes as a screening test for the TYS by using a mixed-methods approach. It will involve the test development stage of a Turkish C-test and two validation studies in different contexts. Kane's (2006) argument-based approach is used to inform the validation studies because it is pragmatic and starts the validation process from the perspective of what the test is to be used for rather than merely focusing on its statistical properties in isolation (see section 2.4.4 for details about the argument-based approach). In laymen's terms, an argument-based approach links candidates' test performance to test uses through a range of test assumptions which are evaluated during the validation process. It gives guidance to lay out the assumptions based on test scores and justify the assumptions through a framework of inferences (scoring, generalization, extrapolation, and decision). Furthermore, an argument-based approach provides enough flexibility to be used in any kind of test regardless of factors such as the content of the test and the stakes of the test since test validation process depends on test uses and interpretations (see Gaillard (2014), Drackert (2016), and Pardo-Ballester (2010) for the examples of using argument-based approach in the SLA context and L2 classroom).

1.4 Outline of the Dissertation

Following this Introduction Chapter, Chapter 2 gives theoretical background about models of L2 proficiency and then describes how L2 proficiency is conceptualized in this dissertation. Then, it presents C-tests as a short-cut measure of L2 proficiency.

Finally, it reviews the literature on validity and validation approaches with an emphasis on Kane's (2006) argument-based validation approach.

Chapter 3 situates the assessment of L2 proficiency in the Turkish context by providing information about the status of Turkish as an L2 including Turkish SLA research and the development of the TYS. Then, before moving on to the Turkish C-test, morphology of the Turkish language (the study of the word structure and word formation) is discussed in terms of the challenges it can create in the development of a Turkish C-test as well as the ways researchers responded to these challenges.

Chapter 4 presents methodological choices used across the two validation studies (Chapter 6 and Chapter 7) as well as the initial investigation of the test development (Chapter 5). In order to justify the methodological choices, it describes the underlying epistemology and the philosophical assumptions of the dissertation. Then, data collection procedures common to both validation studies and background information about the data analysis methods are explained.

Chapter 5 presents the steps of the development of the Turkish C-test. Then, it presents the initial investigation with Turkish L2 speakers. It gives background information about participants, instruments, data collection and data analysis procedures. Following this, it presents the findings, and discusses how these findings guide the next validation studies.

Chapters 6 and 7 present the two validation studies that evaluate two different uses of the Turkish C-test in SLA research and educational assessment. Both validation studies start by introducing the interpretive arguments developed for each test use and then the evaluation questions generated from the interpretive arguments. The methodology involving participants, instruments, data collection and data

analysis procedures are detailed. Following this, results of the analysis for each evaluation question are presented and discussed.

Chapter 8 provides a general discussion and summary of findings across test development and two validation studies focusing on the wider relevance of this research for both the academic literature and practice. Then, it addresses the limitations and provides suggestions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

There are two main questions to bear in mind in validating any kind of test: (1) what is the theoretical model underlying your test? (2) how do you generate evidence concerning the model in practice? (O'Sullivan & Weir, 2011). To explain why the specific test (C-test) and theoretical model of validation (Kane's 2006 argument-based approach) have been chosen in this study, the literature review will first focus on the construct of L2 proficiency and how it is assessed in SLA research and educational contexts. Then, it will introduce cloze tests and C-tests as an alternative way to measure L2 proficiency. Finally, it will give an overview of current validation methods in language assessment.

2.2 L2 Proficiency

Defining the construct of L2 proficiency is a very challenging part of validating a test as an L2 proficiency instrument (Hulstijn, 2012, 2015; Norris & Ortega, 2012). There exist different theories about the construct of L2 proficiency. These theories and their operationalization in SLA as well as educational assessment will be explained in this section of the chapter in order to explain how L2 proficiency is conceptualised in this research.

2.2.1 Models of L2 Proficiency

Several L2 proficiency models have been proposed to define language ability (for a review, see McNamara, 1996; Chalhoub-Deville, 1997; Bachman & Palmer, 1996). Earlier models of L2 proficiency were two dimensional; one dimension involved linguistic knowledge (i.e., grammar, lexicon, pragmatic), and the other dimension consisted of four language skills (reading, writing, speaking, and listening) in which

the elements of the linguistic knowledge were integrated (see Carroll, 1961; Lado, 1961).

In contrast to the dimensional models of L2 proficiency, Oller (1979) proposed the concept of 'unitary' language model, and he claimed that language proficiency involved a global language factor rather than different components. He suggested to measure the global language factor through integrative tests (Carroll, 1961) such as C-tests and dictation tests which require test takers to combine different elements of linguistic knowledge in completing a task. For example, completing the gaps in a C-test requires test takers to combine their grammatical and lexicon knowledge in an embedded context.

Oller's unitary language model was criticised by other scholars based on further models showing the multicomponential structure of language proficiency. These models involved communicative competence model and communicative language ability (CLA) model. Communicative competence model (Canale & Swain, 1980) consisted of three elements: grammatical (lexicon, morphology, syntax, phonology, semantics), socio-cultural (i.e. rules of discourse), and strategical competence (i.e., communication strategies). It was pioneering in introducing the communicative competence into language teaching and testing; however, it did not specify how the elements of language proficiency interact with each other in language use. Addressing the limitations of communicative competence model, CLA model (Bachman, 1990) attempted to include the interaction between various elements of language use. Later, Bachman and Palmer (1996) elaborated on the CLA model and formed a three-level hierarchical model of language ability as shown in Figure 1.

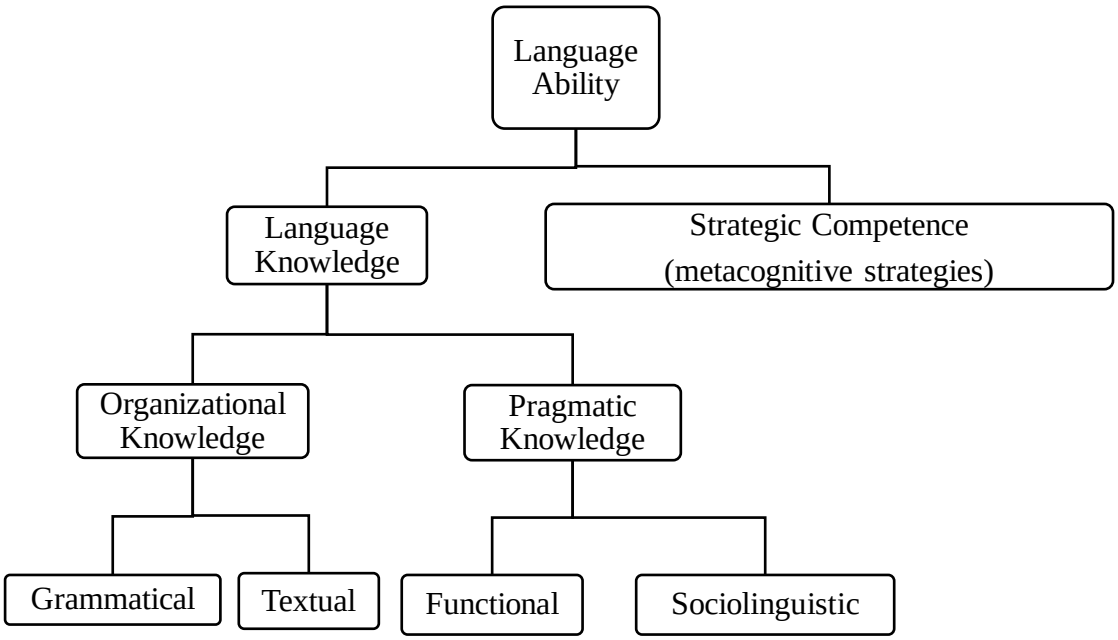


Figure 1. Communicative Language Ability Model (Hulstijn's (2015, p. 39) adaptation of Bachman and Palmer (1996, p. 66-68))

As seen in Figure 1, language ability includes not only purely linguistic components, but also non-linguistics components (strategic competence) which are resorted to in language use. Language knowledge is divided into two sections: (1) organizational knowledge, which relates to how individuals produce grammatically correct language, and (2) pragmatic knowledge, which relates to how individuals produce contextually appropriate language (Purpura, 2004). Organizational knowledge is further divided into two components: (1) grammatical knowledge involving vocabulary, syntax and phonology/graphology, and (2) textual knowledge comprising cohesion, rhetorical organization or conversational organization. Pragmatic knowledge is divided into two categories as functional knowledge (i.e., how to use organizational knowledge for communicative goals) and sociolinguistic knowledge (i.e., how to use organizational knowledge in accordance with language-use context).

Based on these models of language proficiency, the multicomponential structure of language ability has been empirically supported; however, it remains uncertain what the components of language proficiency are and how they interact with each other in language use (Douglas, 2000; O'Sullivan & Weir, 2011; Purpura, 2008). For example, O'Sullivan and Weir (2011) found the CLA model difficult to operationalise in language testing because it wasn't clear which components form the criterial elements for proficiency assessment. They commented that the cognitive processing dimensions of CLA model is not adequate to differentiate between different levels of proficiency and to be used for test development purposes. The field of language testing still lacks a consensus language proficiency model. According to Bachman (1990), this is because the context in which language is used affects the underlying ability.

Although there is not a consensus L2 proficiency model, all models of L2 proficiency agree on the general components of language proficiency, which are grammar and lexis. Hulstijn (2015) found that knowledge of grammar and lexis were strongly associated with performance in four main skills tests. He claims that this result forms a preliminary support for the core-periphery distinction in his model of L2 proficiency as shown in Figure 2 below. In this model, core elements are inclusive of but not limited to grammar and lexis as explained below.

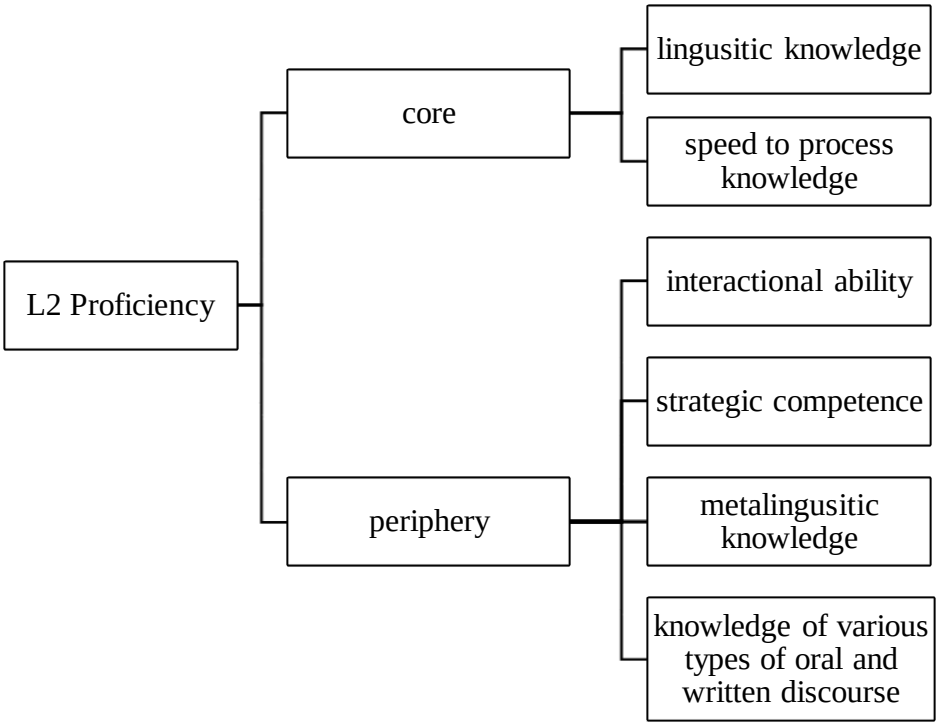


Figure 2. Core and peripheral components of L2 proficiency (adapted from Hulstijn, 2015, p. 42)

Hulstijn's (2007, 2012, 2015) language proficiency model is the first model of L2 proficiency that makes the core-periphery distinction regarding various proficiency components and combines linguistic knowledge with the speed to process (i.e., produce or perceive) knowledge as the core component of language proficiency. According to this model, linguistic knowledge involves not only grammatical and lexical knowledge, but also, knowledge about which language to use in different contexts (pragmatic, sociolinguistic and discourse knowledge). While the knowledge of vocabulary and grammar (and pronunciation) is in the core of language proficiency and distinguished from non-linguistic components at the periphery level, it is not stated how independent these core components are from each other or whether one component is claimed to have a bigger effect on one of the integrated skills (combination of two or more language skills such as listening and speaking in a dialogue). The empirical claim of this core-periphery distinction in Hulstijn's theory

is that core tests will significantly correlate with four main skills tests while peripheral-component tests will correlate with only some tasks in skills tests, which is tested in dissertation by investigating the correlation between a core test (C-test) and four-main skills test (TYS).

Hulstijn (2015) recommends using level-appropriate tests of core skills as a complement to tests of integrated skills. For example, a single writing test on one topic would not be sufficient to assess writing skill at B2 level as it is stated in Common European Framework (CEF): "Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources" (Council of Europe, 2001, p. 61). Therefore, depending on the test purpose, a complementary test of vocabulary, grammar, and spelling at B2 level of CEF should be administered. Although it is not done in this thesis due to the low stakes nature of the suggested test uses and the need for practicality for these low stakes uses, Turkish C-test might be used as a complementary component skill test to an integrated-skills based Turkish proficiency test in higher stakes contexts since it requires knowledge of grammar and lexis in an embedded context. This suggestion is in line with recent research that supports reporting both an overall score and scores for four different skills (i.e., Sawaki, Stricker & Oranje, 2009; Stricker, Rock & Lee, 2005).

In addition to the core and periphery components of L2 proficiency, Hulstijn (2011, 2015) also differentiated between basic and higher language cognition (BLC and HLC) where cognition refers to a neural network comprising both representation and use of language knowledge. According to him, BLC is the common language cognition among all native speakers regardless of varying factors such as age and literacy level, and it involves frequently used lexical items and grammatical

structures. BLC is restricted to speech production and speech perception. HLC is complement or extension of BLC and comprise language cognition which might differ among native speakers. HLC is not essential for native speakers to acquire and it may differ enormously in each individual native speaker while BLC is acquired by all native speakers. Therefore, there is not an ideal and single version of 'the' native speaker except the sharing of BLC. BLC and HLC distinction is important for this dissertation since it guided the decision on recruiting native speakers with a certain level of literacy and educational background (i.e., adults who have completed at least high school education) for piloting the Turkish C-test.

2.2.1.1 Defining L2 proficiency in the present research

Based on the models of L2 proficiency described in the previous section, the current consensus in language testing is that language proficiency is multicomponential, and it comprises a general language factor as well as specific factors. Thus, the construct of language proficiency can be construed as both unitary and divisible depending on test purpose (Harsch, 2014). The present research construes language proficiency as a unitary concept since it aims to provide short-cut estimates of general language proficiency with a single score. Carroll (1993) defined general factors of language proficiency as the measures of a learner's knowledge of grammar and lexicon and progress of language development. The C-test involves these general factors and can cover a broad spectrum of language development by involving texts of different levels according to different frameworks (i.e., CEFR, ILR, ACTFL). It is a core test (Hulstijn, 2012, 2015) that requires knowledge of both grammar and lexis in an embedded context, and it is expected to correlate with four main skills tests. Therefore, although the C-test doesn't directly measure these four skills, it can be used as predictive of proficiency in skills tests, which will be tested in Chapter 7.

2.2.2 Assessment of L2 Proficiency in SLA Research

Despite the variety of models to define language proficiency as explained in section 2.2.1, defining and measuring language proficiency has been underestimated in SLA research (Hulstijn, 2012; Norris & Ortega, 2003, 2012).

Thomas (1994) grouped proficiency measurements used in SLA research under the following four categories: impressionistic judgments (teachers' or researchers' opinion of the test taker's proficiency), institutional status (the test taker's year or level of language study), in-house assessment (locally designed proficiency tests), and self-reported standardized test scores. Among these methods, institutional status was the most common, followed by standardized tests, impressionistic judgement, and in-house assessment.

Following this, Tremblay (2011) found that most SLA studies did not measure L2 proficiency with an independent measure of proficiency. She also observed that some studies solely depended on institutional status and year of study to determine L2 proficiency. However, these indicators of L2 proficiency are not systematic and replicable, and they don't provide generalization and replication across studies. This is because homogeneity does not exist between curricular levels of different institutions and even within the same classroom. Therefore, using an independent proficiency instrument such as C-test would provide more systematic and replicable results rather than crude proxies such as the year of study.

Following Tremblay (2011), Hulstijn (2012) reviewed a corpus of 140 empirical papers and looked at how language proficiency was measured. Similar to Tremblay (2011), he found that language proficiency was not measured with an independent proficiency instrument in 55% of the articles where the measurement of language proficiency was necessary as an independent variable. The effect of

language proficiency was not used to explain the variance in dependent variables, either. This situation becomes more common in LCTL due to the lack of standardized language proficiency tests publicly available to researchers. For example, Montrul (1997) could not find a standardized test of Turkish or Spanish proficiency when she needed an independent measure of L2 proficiency to compare L2 acquisition of verb classes by intermediate level learners across three different languages: English, Spanish, and Turkish. Therefore, she used the Turkish and Spanish translations of an English cloze test that was created based on a passage in an American advanced level course book. However, she found out that the Turkish and Spanish control groups (with native speakers) performed barely above 50% accuracy rate while it was 67% for the English control group probably due to the fact the test was originally English. As Hulstijn (2015) commented, languages differ in the number and nature of their grammatical and lexical elements. Therefore, tests which aim to assess the same linguistic knowledge might measure different things in different languages. Learners of different languages getting the same score on the same test translated into different languages does not mean that their level of language proficiency is the same.

Overall, in SLA research, there is a need for "short-cut" measurements that can provide reliable estimates of language proficiency within researchers' time limitations and enable generalization, replicability and interpretability across studies.

2.2.3 Assessment of L2 Proficiency in Educational Contexts

Educational assessment may come in a large variety of forms from high-stakes and norm-referenced national assessments to low-stakes and classroom-based diagnostic assessments based on test "purposes, uses, users, and contexts" (Norris, 2008, p. 1).

While there is a large body of research on high stakes performance-based language proficiency tests that are used for gate-keeping purposes such as admission

and enrolment of international students into a university, there is relatively little research on diagnostic and screening tests that are used to provide information to teachers and learners (Alderson, et al, 2015; Elder, & von Randow, 2008; Schmidgall et al, 2017). Nevertheless, these tests can have a powerful impact influencing teaching practices or learners' views of themselves. Therefore, they are also significant to facilitate decision making from the point of learners and teachers.

Schmidgall et al (2017) grouped the purposes of educational screening tests under two categories: (1) identifying learners as a member of a particular learner population (i.e., Abedi, 2008; Bailey & Kelly, 2011; Mahoney & MacSwan, 2005); (2) classifying learners as a member of a more or less proficient group or according to the sufficiency of their levels to take a particular exam (i.e., Stansfield & Hewitt, 2005; Xi, 2008). Hence, screening tests can be useful to help candidates to decide whether a high-stakes exam is appropriate for their level, and thus, they can prevent wasting time and money for an exam unsuitable for their level. Considering their use in directing candidates to higher stakes exams, screening tests should be practical to administer, score, and complete within a short amount of time (Schmidgall et al, 2017). Validating the uses of these tests are essential when they are published online open to thousands of learners to provide them immediate information about their levels (Chapelle et al, 2003).

2.3 Short-cut Estimates of Language Proficiency

In response to the lack of validated short-cut estimates of language proficiency, there have been studies to validate alternative measurements of language proficiency such as cloze tests and C-tests in various languages (i.e., Gaillard, 2014; Drackert, 2016; Norris, 2018; Tremblay, 2011). As Norris (2018, p.11) explained, 'short-cut' estimates are defined as "the variety of language assessments that can, within

relatively few items and short test-administration time, provide reliable and accurate estimations of holistic language proficiency across a broad range of levels". Reduced redundancy test types are fit as the short-cut estimates of language proficiency since they can be developed, administered, and scored in a short amount of time while they involve integrative skills requiring test takers to combine different elements of linguistic knowledge at the same time. On the other hand, commonly used test types such as multiple-choice tests are relatively more difficult to create (i.e., creating sensible distractors) and involve guessing factor while not necessarily involving integrative skills. Thus, they have not been chosen as an assessment tool for the SLA research and educational screening purposes of this dissertation.

The C-test is the chosen measurement technique in this since it is a type of reduced redundancy test and deemed to be a relatively under-researched but promising area in the assessment of LCTL compared to other alternatives. Furthermore, it has several advantages over cloze tests which will be explained in section 2.3.2.

Since C-tests were originated from cloze tests and introduced as an alternative to the shortcomings of cloze tests, first cloze tests will be briefly explained, and then C-tests will be detailed in the rest of this section.

2.3.1 Cloze Tests

Taylor (1953) introduced the 'cloze principle' to language studies claiming that it measures the readability level of texts. He traced the cloze principle to the Gestalt psychology and information theory, in that, people can still process information when part of the information is missing. Regarding languages, readers can supply missing linguistics items even when texts are altered because natural languages are redundant to some extent, that is, the same information is expressed more than once in different

ways within a sentence (Spolsky, 1969, 1973). Based on this reduced redundancy principle of languages, Taylor developed cloze tests by deleting some words in a text. Readers were expected to restore deleted words in texts depending on their reading ability.

Taylor also suggested to use cloze tests to compare native speakers' reading abilities since he found differences between readers' scores. Therefore, cloze tests started to be used to compare first language (L1) speakers' reading comprehension ability. In the early 1960s, cloze principle was introduced to second language (L2) testing as a measure of general language proficiency. Although reading factor had a determining impact on test scores, cloze tests were rather considered as measures of general language proficiency due to restricted content coverage and cognitive processing involved compared to reading tests.

Oller and Conrad (1971) did a study about the discriminative power of cloze tests, and they found that the cloze test discriminated well between beginning, intermediate and advanced level English as a second language (ESL) students, but it failed to distinguish between advanced level ESL students and English L1-speaking freshmen. Similarly, Tremblay (2011) showed that the French cloze test was not able to distinguish among very high-level and very low-level learners despite its overall good results and its power to place L2 learners into different levels. These findings might be attributed to that cloze test measures proficiency in written modality and might not closely reflect L2 learners' aural/oral proficiency. Written and spoken proficiency should correlate with each other due to the general language factor explained earlier, but they may not be exactly alike. Therefore, reduced redundancy tests in written modality can be administered to L2 learners together with the ones in spoken modality depending on intended uses of the test. Reduced redundancy

principle was also used in spoken modality including dictation test, the noise test, and partial dictation test in addition to a variety of cloze tests in written modality such as multiple-choice cloze test and rational deletion cloze test (Sigott, 2004).

2.3.1.1 Limitations of Cloze Tests

Alderson (1978, 1979) suggested that cloze tests prompt test takers only to look at the immediate words surrounding the gap to complete it. Therefore, he claimed that cloze tests involve primarily lower-order language skills and don't often measure text-level processing. Klein-Braley (1981) supported Alderson's claim that cloze tests may only test lower-level skills. She could not find inter-dependency between items in a cloze test, which would ensure using textual knowledge (i.e., cohesion, rhetorical organization) while completing the gaps. Klein-Braley and Raatz (1984) summarised the main practical and theoretical reasons why cloze tests are not sufficient as follows: (1) systematic nth word deletion does not necessarily produce a random sample of the elements of the text, (2) different deletion rates and starting points produce tests with different levels of difficulty, (3) text selection is difficult considering its suitability for the sample, (4) there is a high chance of test bias since examinees are presented with only one or two texts. Baghaei and Grotjahn (2014) added that the application of Cronbach's Alpha formulas and Rasch analysis might be problematic in cloze tests due to the assumption of local independence of items.

2.3.2 C-tests

Addressing the limitations and criticisms of cloze tests, Klein-Braley and Raatz (1982) suggested a modified version of the cloze test called the C-test where the 'C' stands for cloze. They introduced the C-test as a better representation of the reduced-redundancy principle (Klein-Braley & Raatz, 1982; Raatz & Klein-Braley, 1982). In C-tests, several shorter texts involving a larger number of items can be completed

within a shorter amount of time while cloze tests usually consist of one or two longer texts with less items. For example, a total of 100 to 125 items (i.e., 20 to 25 gaps per text) can be completed in a 5-text C-test within half an hour. On the other hand, rather long texts are required to have approximately 25 items in one cloze test text, and using only one or two text due to the text length could cause a potential context bias. Therefore, C-test are more practical for the quick research and screening purposes of this dissertation. Furthermore, C-tests require a greater level and variety of linguistic knowledge since learners are expected to reconstruct every second word rather than predicting or choosing the most suitable every sixth or seventh word as in a cloze test (Norris, 2018). In the rest of this section, first, the structure (format) of C-tests will be explained, and then, the construct and appraisal of C-tests will be discussed.

2.3.2.1 The Structure of C-tests

According to Raatz and Klein Braley (1985), the first step in developing a C-test is that four to six tests with around 60-70 words are chosen. However, Grotjahn (1987) suggested that a researcher should begin with at least twice as many C-test texts as the actual test will consist of, as some of the texts may be excluded due to statistical properties and other factors (i.e., low accuracy percentage with native speakers). Ideally, the texts are selected from authentic resources (Klein-Braley, 1997). These texts should be neutral in content, appropriate to the target group, and without any requirement for highly domain-specific vocabulary or knowledge. As Norris (2006) noted, overly technical, bizarre, or infrequent texts should be avoided, as should texts with extensive use of proper nouns. The first sentence in each text is left intact to give the test taker a general understanding about the content of the test. Then, the second half of each second word, starting from the second word in the second sentence, is deleted. If words have an odd number of letters, the second half of the word plus one

letter is deleted. Acronyms, proper nouns, one-letter words, numbers, and dates written numerically are left intact. After the 20th or 25th deletion in a text, words are no longer deleted. To allow the text to come to a natural end, the last sentence is left intact (see an example from Dörnyei & Katona, 1992, p.205 below).

> One cool autumn evening, Bob L., a young professional, returned home from a trip to the supermarket to find his computer gone. Gone! all so____ of cr____ thoughts ra____ through h____ mind: H____ it be____ stolen? H____ it be____ kidnapped? H____ searched h____ house f____ a cl____ until h____ noticed a sm____ piece o____ printout pa____ stuck un____ a mag____ on h____ refrigerator do____. His heart sank as he read this single message: CAN'T CONTINUE, FILE CLOSED, BYE. [7]

Once the test is developed, the test should be piloted with a control group of adult native speakers with a certain level of literacy. It is scored by assigning one point to each fully correct answer. Scores for each text (20 or 25 points) and the whole test (120 or 150 points) are calculated. The acceptable level of accuracy expected from pilot testing with native speakers is around 95% on average (Klein-Braley, 1985). During pilot testing with native speakers, if possible alternative solutions are found for deleted words, they might be included as permissible answers (Klein-Braley & Raatz, 1984). Texts which native speakers cannot complete with around 95% accuracy should be replaced with new texts. After pilot testing with native speakers, the test is administered to a large group of L2 learners. The facility and discrimination indices of each text are calculated and the least satisfactory texts (i.e. too easy, too difficult) are discarded. Before or after the removal of the unsatisfactory texts, the overall test reliability is calculated through Cronbach's Alpha.

[7] Solutions: *sorts, crazy, raced, his, had, been, had, been, he, his, for, clue, he, small, of, paper, under, magnet, his, door.*

2.3.2.2 What do C-tests measure?

The way learners perform under the conditions of C-test deletion is believed to provide evidence of their general language proficiency. The more proficient a learner is, the less redundancy s/he will require for effective performance, thus, observing how a learner deals with reduced redundancy seems to be a great way of determining his/her general language proficiency (Sigott, 2004). Therefore, C-tests have been used to provide a general estimate of language proficiency both in SLA assessment and educational assessment such as placement and screening tests (e.g., Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Eckes, 2014; Lee-Ellis, 2009; Norris, 2006, 2018). As discussed in section 2.2.1.1, the construct of general language proficiency is inclusive but not limited to grammar and lexis. C-tests require knowledge of grammar and lexis; however, they also tap into the other elements of language proficiency depending on text difficulty and learner proficiency.

According to Harsch and Hartig (2016) the C-test measures language knowledge and skills as well as metacognitive strategies. Thus, it requires integrating "contextual, semantic, syntactic, morphological, lexical, and orthographic information and knowledge" (Hastings, 2002, p. 66). Since C-tests are integrative tests, they can measure the global language proficiency as suggested by Oller (1979). This is in line with the operationalization of language proficiency as a unitary concept in this dissertation (see section 2.2.1.1).

Sigott (2004) argued that the construct measured by the C-test is "fluid" since he found that the amount of context required to solve a C-test item depends on individual test takers' language proficiency. While some test takers use the whole passage at the text level to solve a C-test item, more proficient learners can operate with more limited context (i.e. word level, sentence level) to solve the same item. It

seems that while the C-test measures the general factors of grammar and lexis, which are shared by all models of L2 proficiency, it taps specific factors of L2 proficiency only in some learners depending on the learners' language proficiency and the level of text difficulty. Less proficient learners depend on more context to compensate for their lack of knowledge. Therefore, the aspects of the construct which are tapped by the C-test depends on the test taker proficiency as well as text difficulty.

C-test difficulty depends on macro level factors (i.e., the inter-gap dependency, the paragraph difficulty) as well as micro level factors (i.e., solution difficulty, the candidate ambiguity) (Beinborn, Zesch & Gurevych, 2014). Evidence for this comes from Khoshdel, Baghaei and Bemani (2016) who found that gap-level (word and sentence level) factors explained for only 8% of variance in text difficulty leaving 92% variance open for factors beyond the lower level word and sentential ones. Therefore, they recommend researchers to deal with the effect of text characteristics on test difficulty. It is, yet, unclear whether even the most difficult C-tests can distinguish well between high-level proficient learners and what the relationship between components of C-test construct is.

Despite the integrative nature of C-tests, there are some studies which consider C-tests as reading or vocabulary tests (i.e., Chapelle, 1994; Cohen, Segal, & Weiss, 1985; Read, 2000). However, C-tests don't comprise all complex levels of cognitive processing involved in reading tests (see, for example, Khalifa & Weir (2009)'s model of reading to examine levels of cognitive processing). C-tests also involve limited productive skills which are not necessarily required in reading comprehension ability. Furthermore, they cannot be reduced to a vocabulary test since they require textual understanding due to their embedded and contextualized structure depending on text difficulty.

2.3.2.3 Appraisal of C-tests

C-tests are usually applauded for their high reliability and validity indices, the ease of test construction, administration, and scoring compared to other standardized tests of L2 proficiency (Klein-Braley & Raatz, 1984; Klein-Braley, 1997; Eckes & Grotjahn, 2006). Although it was initially claimed that C-tests measure only micro-level processing at the word and sentence level, they were later shown to measure both micro and macro-level processing depending on examinee proficiency level and contextual factors (i.e., Babaii & Ansary, 2001; Feldmann & Stemmer, 1987; Grotjahn, 2002; Klein-Braley, 1994; Sigott, 2004).

C-tests strongly correlate with the four main skills tests and lexis tests aligning with Hulstijn's (2015) language proficiency model (see section 2.2.1). Sigott (2004) reported a detailed analysis of the criterion-related validity of C-tests conducted between 1990 and 2002. Following this, Eckes and Grotjahn (2006) examined the correlational studies of C-tests and various language tests conducted between 1987 and 2006. The criterion tests were well-established tests including TOEFL, TestDaF, TOEIC, English Language Battery (ELBA), and several locally used department and placement tests. Both studies found that the correlations were generally quite high between C-tests and total scores in criterion tests such as TOEFL and placement tests. There were also moderate-to-high correlations between C-tests and four different skill tests as well as grammar and vocabulary tests. These correlational studies are shown in Table 1 below which was adapted from and expanded on Sigott (2004) and Eckes and Grotjahn (2006).

Table 1. Correlations between C-tests and other language tests

Study	Language	Subjects	criterion	Reliability
Eckes, 2014	6 German C-tests with 8-texts	pre-university students of L2 German (N=1,467)	TestDaF reading; *r* = .61 to .73 listening; *r* = .63 to .82	.92 to .95
Arras, Eckes, & Grotjahn, 2002	German C-test with 4 texts	university students of L2 German (N between 145 and 187)	TestDaF reading; *r* = .65 listening; *r* = .64 writing; *r* = .68 speaking; *r* = .64	.84
Babaii, Ansary, 2001	English C-test with 5 texts	Iranian EFL students (N=32)	TOEFL total; *r* = .88 structure; *r* = .88 vocabulary; *r* = .79 reading; *r* = .80	.88
Daller, Phelan, 2006	English C-test with 6 texts	French EFL students (N=30)	TOEIC reading; *r* = .48 listening; *r* = .45	.84
Jafarpur, 2002	English C-test with 4 texts	Iranian EFL students (N=146)	English Placement Test total; *r* = .87 reading; *r* = .86 listening; *r* = .87 grammar; *r* = .84 vocabulary; *r* = .85 cloze; *r* = .78 (correlations corrected for attenuation)	.92

Chapelle, Abraham, 1990	English C-test with 5 paragraphs from one text	university students of L2 English (N=49)	English Placement Test reading; *r* = .60 listening; *r* = .47 writing; *r* = .64 vocabulary; *r* = .84 (correlations corrected for attenuation)	.81
Dörnyei, Katona, 1992	English C-test with 4 texts	Hungarian university English majors (N=102)	TOEIC total; *r* = .62 reading; *r* = .54 listening; *r* = .51 Oral Interview; *r* = .43 Department Proficiency Test total; *r* = .43 listening; *r* = .33 vocabulary; *r* = .38 grammar; *r* = .25 Cloze Test; *r* = .38	.75
Grotjahn, Allner, 1996	German C-test with 8 texts	pre-university students of L2 German (N=141)	Language Admission Exam grammar; *r* = .75 oral reproduction test; *r* = .81	
Negishi, 1987	English C-test with 4 texts	Japanese EFL university students (N=20)	ELBA total; *r* = .76 reading; *r* = .80 grammar; *r* = .56 vocabulary; *r* = .62	.78
Boonsathorn, 1987	2 English C-tests with 4 texts	ESL university students (N=23, 19)	Michigan Test total 1; *r* = .54 total 2; *r* = .61	.81 .90
Sigott, 2004	English	Austrian students of English	Oxford Placement Test	.81

	C-test with 4 texts	(N=60)	grammar; *r* = .97 (correlation corrected for attenuation)	
Chihara et al, 1996	4 English C-tests with 4 texts	Japanese EFL university students (N=82 to 93)	TOEFL total; *r* = 55 to 65 structure; *r* = 43to 61 vocabulary & reading; *r* = 38 to 50 listening; *r* = 36 to 61	.76 to .81
Harsch, Hartig, 2016	English C-test with 4 texts	German secondary school students (N=559)	Large-scale German exam reading; *r* = 73 listening; *r* = 76	.95
Norris, 2006	2 German C-tests with 5 texts	university students of L2 German (N= 92 to 102)	Department Placement Test reading; *r* = .84 to .86 listening; *r* = .77 to .82	.95 to .96
Drackert, 2016	Russian C-test with 4 texts	university students of L2 Russian (N=67)	EIT *r* = .79	.90
Son, 2018	Korean C-test with 4 texts	university students of L2 Korean (N=93)	EIT ρ = .84 ACTFL writing; *r* = . 85 ; ρ = .87 OPI; ρ = .81 writing + OPI; ρ = .80	.94

As seen in Table 1, C-tests have moderate to high correlations with all language tests, which shows that they gauge several skills simultaneously and holistically. The only correlation that appeared to be relatively small (.25) is between the C-test and a grammar test (Dörnyei & Katona, 1992). Dörnyei and Katona (1992) attributed this to word-level deletions rather than sentence-level deletions in C-tests. Also, it is acknowledged that C-tests correlate with tasks which require higher level skills more than they do with discrete-point grammar and vocabulary tests (Harsch & Hartig, 2016). Harsch and Hartig (2016) correlated C-test and Yes/No Vocabulary Test, both of which are used as screening and placement tests, with listening and reading comprehension (receptive skills). They found that C-test correlated with reading and listening comprehension skills (r=.73 and r=.76 respectively) higher than Yes/No Vocabulary Test correlated with these skills (r=.39 and r=.49 respectively) due to its embedded and contextualized structure.

Although the correlations between C-tests and oral productive skills tests are medium to high, there have been only a few correlational studies of this type as can be seen in Table 1 and more studies are needed, which will be addressed in Chapter 7 by correlating the Turkish C-test with the TYS speaking section. Dörnyei and Katona (1992) reported a correlation coefficient of .43 between an English C-test and oral interview. Grotjahn and Allner (1996) did a factor analysis of C-test with a task requiring the reproduction of an orally presented test. They found that C-test had a high correlation of .81 with the text reproduction task. Furthermore, the correlations between C-tests and speaking sections of TestDaF were high ranging from .64 to .54 (Arras et al, 2002; Eckes, 2014). Recently, Son (2018) looked at the correlations between a Korean C-Test and ACTFL (American Council on the Teaching of Foreign Languages) OPI (Oral Proficiency Interview) as well as EIT. She found that C-test

correlated with EIT (ρ=84) slightly higher than it correlated with ACTFL OPI (ρ=81) probably because EIT is short-cut measure of language proficiency based on reduced redundancy principle as C-tests where learners have to comprehend and process the holistic meaning of sentences. Therefore, although the C-test does not directly measure oral/aural skills, it correlates with skills tests since grammar and lexis are common to both.

In addition to their correlations with criterion tests, C-tests have also been shown to have moderate to high correlations with program levels (i.e., level of the course being taken) and self-assessments of proficiency. Recently, Norris (2018) investigated these correlations in various languages as seen on Table 2.

Table 2. Correlations between C-tests and program level and self-assessment

5-Text C-Test	Program level	self-overall	self-reading	self-speaking	self-listening	self-writing
French	.85	.63	.67	.58	.67	.59
Japanese	.84	.79	.68	.62	.69	.66
Arabic	--	.63	.76	.64	.41	.66
Portuguese	.58	.65	--	--	--	--
Korean	.79	.78	.68	.77	.69	.68
Bangla	--	.41	--	--	--	--

Regarding the perception of C-tests, there are only few systematic studies, and these studies show that C-tests, in general, lack face validity from the point of learners and teachers (Sigott, 2004). For example, Legenhausen (1989) found that teachers' criticism of the C-tests centred around extreme difficulty level, lack of authenticity, and student frustration by the C-test format. Nevertheless, the actual student scores showed that C-tests were generally too easy for students in contrast to teachers' opinions. Legenhausen commented that this discrepancy might be due to students' and teachers' lack of familiarity with the test format and the random selection of texts that might include vocabulary or grammar structure that was not covered in the class.

He went on to say that while students might have expected to show a high degree of accuracy as in a criterion-referenced achievement test, test developers designed the C-test as a norm-referenced test where there was no expectation for a high level of accuracy, and the test involved items with differing difficulty level. Along similar lines, McBeath (1989) said that test takers' initial reaction to C-tests might be "bewilderment" due to the lack of authenticity of the test format and test takers' lack of familiarity with C-tests since they have limited uses in mainstream testing.

Following this, Huhta (1996) did not find a significant correlation between the students' ratings of the face validity of the C-test (i.e., how good the C-test is as a measure of their language proficiency) and their actual scores on the C-test. Furthermore, he said that while some students liked the C-test, many did not, which was because they felt that what the C-test measured was not clear and not necessarily within the construct of language proficiency such as imagination and inferencing. More recently, Sumbling et al (2014) found that the feedback from test takers and teachers indicated lack of reliability which contradicted the psychometric evidence.

Overall, while C-tests have been shown to be practical, function reliably and distinguish between learners of different proficiency levels, they have low face validity due to factors such as seeming lack of authenticity, unique format, stakeholders' lack of familiarity with this format, and a resulting frustration (Sigott, 2004). Sumbling et al (2014) suggested providing teachers and students with statistical information related to the reliability and validity of C-tests in order to address this limitation. Nevertheless, as Sigott (2004) noted, the number of systematic qualitative studies investigating stakeholders' perception of the C-test is limited. Study 1 and study 2 will address this gap through the qualitative analysis of interviews and open-ended survey answers.

2.4 Validation Approaches

In this dissertation, the term validation is used as a unitary term to refer to the overall process of investigating validity (i.e., whether the test is valid for suggested interpretations and/or uses made based on test scores) and it is not interchangeably used with validity. Chapelle and Voss (2013) analysed a corpus of 123 empirical studies conducted between 1984 and 2011 to investigate validation approaches used in language assessment. As a result, they categorised validation methods under four groups: 1) one question and three validities, 2) evidence gathering, 3) test usefulness, and 4) argument-based approach. This section will introduce the first three validation methods briefly and then explain the argument-based approach more in detail since it is the adopted validation approach in this dissertation.

2.4.1 One Question and Three Validities

This approach relates to the question "does the test measure what it claims to measure?" (Lado, 1961). To answer this question, it centres around three different types of validity: content, criterion, and construct. It was very commonly used in language testing till 1990 (Chapelle & Voss, 2013). However, following Messick's unitary validity framework, it was criticized for showing validity as a property of the test.

2.4.1.1 Criterion Model

Between 1920 and 1950, a test was considered valid for any criterion that it correlated with. Therefore, validity was deemed as the correspondence between test scores and performance on the external criterion.

Criterion model works very well in conditions such as a test is used to predict job performance in the future. However, in some cases, finding a plausible criterion might be difficult. The main deficit of criterion model is reducing the validity to a

correlation between different tests. It is questionable how the criterion is validated in the first place. Even if the criterion is validated with another external measure, that external measure also needs to be correlated with another measurement. Thus, the ambiguity of how to validate a criterion without recourse to another criterion remains a problem of criterion model.

2.4.1.2 Content Model

Content validity is related to whether items of a test in a specific domain reflect the overall level of skill in that domain. For example, using expert judgment is a way of determining whether test items reflect the universe of items in a specific domain. Content validity works very well with tests of specific skills such as achievement test. However, it is problematic when it is used to validate claims about cognitive skills or theoretical constructs (Cronbach, 1971). Messick (1989) argued that content validity supports evidence for "the domain relevance and representativeness of the test instrument" (p. 17), but it still has a limited role because it doesn't provide evidence that interpretations are directly made from test scores. Therefore, evidence from content validity needs to be supported with other evidences to justify decisions made based on test results.

2.4.1.3 Construct Model

Cronbach and Meehl (1955) proposed construct validity as an alternative to criterion and content validity to use in trait measures when there is no adequate criterion or no domain of content to sample. They claimed that a test has construct validity if there is an association linking the empirical relations between observed test scores and the theoretical relations to constructs that the test purports to measure. For example, if a test is said to measure the construct of L2 proficiency, the observable performance (e.g., test scores) is supposed to correspond to hypotheses and predictions derived

from the construct of L2 proficiency. Although Cronbach and Meehl (1955) implied that construct validity was superordinate and unifying, they did not specifically introduce a framework integrating evidences from content and criterion model as well. Therefore, all three models were perceived as three different types of validity and the choice of any of these models was based on the availability of data by the late 1970s (Guion, 1977).

2.4.2 Evidence Gathering

An evidence-gathering approach pertains to the unitary view of validity which was proposed by Messick (1989) and embraced in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). Messick introduced unitary framework for validity by adopting a broader version of construct validity. He proposed validity as one unitary concept instead of several types of validity. His unified conceptualization of validity shifted the focus from the test towards the scores and inferences derived from the test:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment (Messick, 1989, p. 13, italics in original).

Hence, Messick gathered different evidences on different aspects of validity under a unitary validity framework. The trinity view of validity was replaced with construct validity as a unifying concept to which validity evidences from different sources would contribute. Messick included value implications and social consequences of tests into his unitary framework as shown in Table 3:

Table 3. Facets of Validity as a Progressive Matrix (Messick, 1993, p.13)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV+ Relevance / Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

He called his four-facet framework a progressive matrix because a new concept was added to the basic construct validity in each cell. As a result, validity started to be considered with regard to both interpretations and uses of test scores rather than the test per se, and ethical considerations of test uses started to be emphasised. Messick's unitary framework received a great degree of consensus in testing and evaluation, and it was followed by scholars in language testing (i.e., Bachman, 1990; Weir, 2005). However, it did not provide guidance on how to validate test uses and interpretations (Bachman, 2005; Chapelle, Enright & Jamieson, 2008, 2010; Kane, 1990, 1992, 2011; Shepard, 1993). Shepard (1993, 1997, 2016) criticised Messick's matrix to separate "inherently entwined" aspects of validity and require addressing each aspect independently although it was supposed to be a unitary framework (2016, p. 269). Kane (2011) commented that where and how to begin the validation process was not clear. Furthermore, the inclusion of consequences into the unitary concept of validity has been ignored in many validation studies as noted by some researchers (e.g., Brown, 2008; Cizek, Rosenberg, Koons, 2008; Dunlea, 2015).

Overall, with evidence-gathering approach, it might be problematic to decide how much of what kind evidence is needed and how to synthesize different evidences to justify the uses and interpretations of test scores (Kane, 2001, 2002; Shepard, 1993, 2016).

2.4.3 Test Usefulness

The test usefulness approach was introduced to language testing by Bachman and Palmer (1996). They simplified Messick's validity framework in correspondence with

the needs of language assessment. They operationalized validation in six aspects: construct validity, reliability, authenticity, interactiveness, impact, and practicality. Thus, test usefulness approach provided language testers with a practical and clear guidance while keeping the basics in Messick's framework. Argument-based approach is chosen over test usefulness approach in this dissertation since it provides more flexibility to tailor the required validity evidence based on the specific test use and gives more structured guidance to connect different pieces of evidence.

2.4.4 Argument-based Approach

The argument-based approach is a recent trend in validation research in language testing (Chapelle & Voss, 2013). Although it was developed with the purpose of proposing suggestions to problems in validation practice (i.e., selecting and synthesising different sources of validity) and not specific to language testing, it provides researchers with great flexibility to adapt to any kind of test (Kane, 1992, 2006). It differs from the test usefulness approach in that the operationalization of validation depends on the suggested test use rather than pre-set validity aspects. It gives flexibility to form test assumptions tailored for the specific test use. The required types of validity evidence are determined based on these test assumptions and collected to justify them.

The argument-based approach is consistent with the consensus aspects of validity summarised by Cizek (2016) (see section 2.4.5). First, it starts the test validation process from the intended uses and interpretations of test scores rather than the test per se through a useful and structured guidance which was missing in other validation methods. As Kane, Crooks and Cohen (1999) commented a test score or test procedure is not considered valid or invalid itself, and validity inquiry arises only when interpretations are assigned to test scores. Second, it describes validity along a

continuum of inferences which indicate how much of what kind of evidences are needed. Third, it sees validation as an ongoing and critical process rather than a one-time activity. Therefore, changes can be made, and more evidence can be collected during the validation process if necessary. Although argument-based approach does not offer a solution to all the problems in validation approaches discussed in this section (i.e., validating a criterion without recourse to another criterion), it does not reduce validity to a one-time activity and sees it as an ongoing process. Thus, one can see the weakest chain in a validation argument and make adjustments.

According to Kane (2006, 2013), validation involves two kinds of arguments: an interpretation/use argument (IUA) or interpretive argument and a validity argument.

2.4.4.1 Interpretive Argument

The interpretive argument specifies the inferences and assumptions of test uses and test interpretations. It assimilates to a scientific theory in that both present a framework to interpret observed phenomenon and can be evaluated in terms of their plausibility (Kane, 2006, 2016). Kane et al (1999) used the bridge analogy to explain its inferences. The interpretive argument involves three main inferences that are linked through three bridges: (1) scoring, (2) generalization, and (3) extrapolation. Subsequently, Kane (2001, 2004, 2006) extended the bridge analogy to four-bridge formulation by including a fourth inference called decision as seen in Figure 3.

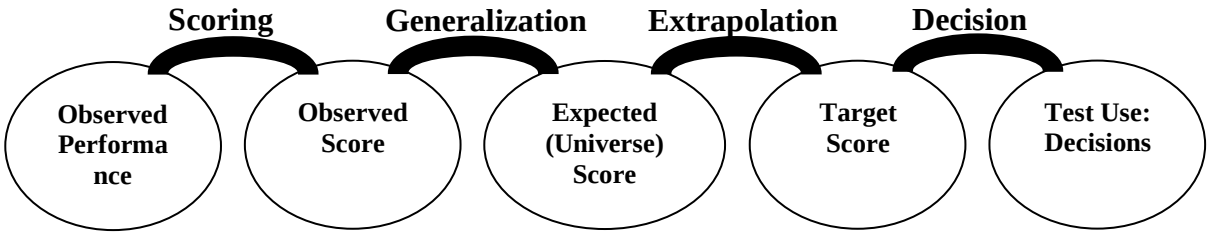


Figure 3. Chain of inferences in argument-based approach (adapted from Kane et al, 1999, p. 9)

The first bridge or inference is scoring, which is assigning a score that to an observed test performance. The second bridge is generalization from the observed score to the universe score that an individual is expected to have over the universe of similar tasks. The third bridge is extrapolation to the target score or actual performance which is the context of what an individual can do outside the test environment in the target domain. Finally, the fourth bridge is decision that is made based on the test score. Through this four-bridge formulation, it is possible to make a decision (i.e., placement, enrolment) based on a test score.

Each of these inferences has their own testable assumptions analogous to the way scientific theories have hypotheses. Kane (2006) exemplified an interpretive argument for a placement test as outlined in Table 4. This model of interpretive argument was modified for the suggested test uses in the research reported in this dissertation (see sections 6.2 and 7.2).

Table 4. Interpretive Argument for a Placement Testing System (Kane 2006, p. 24)

I1: **Scoring**: from observed performance to an observed score
A1: The scoring rule is appropriate
A2: The scoring rule is applied accurately and consistently.
I2: **Generalization**: from observed score to universe score
A1: The observations made in testing are representative of the universe of observation defining the testing procedure.
A2: The sample of observations is large enough to control sampling error
I3: **Extrapolation**: from universe score to the level of skill
A1: The test tasks require the competencies developed in the courses and required in subsequent courses
A2: There are no skill irrelevant sources of variability that would seriously bias the interpretation of scores as measures of level of skill in the competencies
I4: **Decision**: from conclusion about level of skill to placement in a specific course
A1: Performance in courses, beyond the initial course, depends on level of skill in the competencies developed in earlier courses in the sequence
A2: Students with a low level of skill in the prerequisites for a course are not likely to succeed in the course

A3: Students with a high level of skill in the competencies taught in a particular course would not benefit much from taking the course.

Kane (2013) said that flexibility is necessary in developing an interpretive argument since test uses, interpretations, and contexts can be varied. Therefore, he suggested that although some inferences such as scoring and generalization are most likely to occur in all interpretive arguments, there are no strict rules about which inferences to include or exclude. For example, Chapelle et al. (2008) included two more inferences to the chain of inferences and used the following inferences to build a validity argument for the TOEFL: domain definition, evaluation (scoring), generalization, explanation, extrapolation, utilization (decision). In their example, domain definition linked observations of test performance to the real-life performance in target language use domain, and explanation linked expected score to the construct of academic language proficiency. In another example, Son (2018) validated the use of a Korean C-test to measure heritage language learners' proficiency for research purposes, and she preferred the term "theoretical grounds" rather than "domain description" as the first inference since Korean C-test was not a domain-referenced or a skills-test, but a shortcut measure of general language proficiency. Similarly, the validation studies reported in this dissertation modified Kane (2006) by involving the "theoretical grounds" (Son, 2018) as the first inference.

Kane's argument-based approach relies on Toulmin's (1958) model of inference, which was used as a framework to create and analyse arguments in a wide variety of contexts (Kane, 2006). Toulmin's model of inference requires that a claim (i.e., test-score interpretation) be supported and defended if challenged. Toulmin (1958) said that an argument consists of six parts: datum, claim, warrants, backings, qualifiers, and rebuttals. His framework connects a datum (ground) to a claim with

justification provided by a warrant (assumption), which is, in turn, supported by a backing in the form of theories or research as seen in Figure 4. The strength of a claim is determined by a qualifier and warrant does not apply in the case of an exception (rebuttal). Toulmin's model is applied to individual inferences (i.e., scoring, generalization, extrapolation) within an argument. Each inference has its own datum and claim with the conclusion of earlier inferences serving as datum.

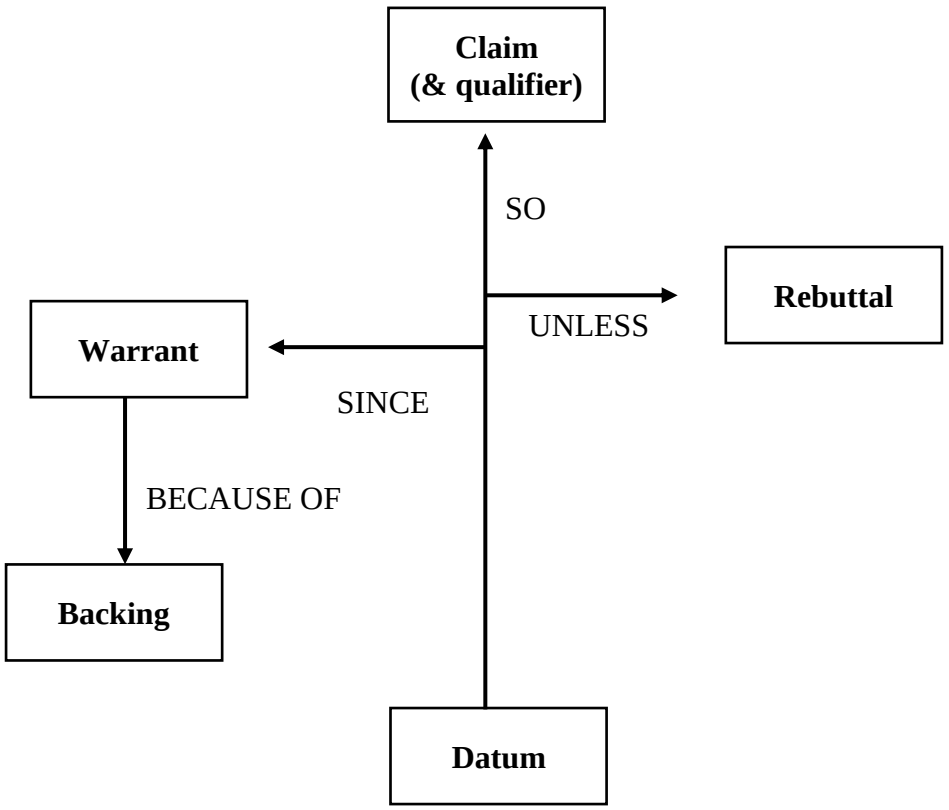


Figure 4. Toulmin's Model of Inference (adapted from Mislevy, Steinberg, & Almond, 2003, p. 11-12)

2.4.4.2 Validity Argument

Kane's (2006) validity argument is the evaluation of the interpretive argument to investigate whether inferences are reasonable, and assumptions are plausible. In other words, the validity argument challenges the interpretive argument by collecting evidence. The kind and amount of evidence depends on inferences and their

underlying assumptions specified in the interpretive argument. Thus, different kinds of evidence are needed to evaluate each inference and assumption involving content analysis, task analysis, expert judgement, empirical studies, correlational analysis, results of previous research, and washback studies (Kane, 2006).

The scoring inference is justified by expert judgement about the appropriateness of the scoring rubric for a test. It is relatively easy to develop scoring rubrics for highly structured assessments such as selected-response tests (i.e., multiple-choice, matching) (Kane et al, 1999). It becomes harder to develop a consistent rubric for complex and open-ended assessments such as performance assessment. When the scoring involves ratings, statistical analyses of agreement among raters (i.e., inter-rater reliability) and the procedures of selecting and training raters may also be required for the validation of scoring inference. If there is too much variation in ratings, necessary changes should be made such as revising the items in question and retraining the raters. When statistical models are used to scale or equate scores, fit of these models to the data should be examined. Another justification regarding scoring inference includes controlling the conditions under which performance is observed. It should be ensured that test administration conditions are suitable for test takers to perform at their level of skill such that they are not exposed to any disadvantages (i.e., inappropriate instructions, technical problems).

The evaluation of generalization requires investigating generalizability of observed scores to universe scores over the universe of generalization, which includes performances over similar tasks. The evidence for this justification is collected through reliability studies (Feldt & Brennan, 1989; Haertel, 2006) or generalizability studies (Brennan, 2001; Chiu, 2001) which provide estimates of standard errors of measurement (how repeated measures of a person's score on the same instrument are

distributed around his/her true score). By doing so, the consistency of scores across samples and the size of the random error is identified. Furthermore, judgments about the representativeness of the sample of observations are used to control sampling error. Any facet (i.e., item, rater) that is subject to variability over the universe of generalization are prone to sampling error. Generalization over items is relatively easy in objective tests, which may include hundreds of test items, compared to performance tests which can include only a small number of tasks due to time and resource limitations and have a potentially larger sampling error. Generalization over raters is also not an issue in objective tests due to highly structured scoring rules while generalization over occasions may be a potential problem. Kane (2006) suggests two options to proceed the validation process if the sampling error resulting from a facet is large: first, the measurement procedure can be changed such as increasing the sample sizes of the facet; second, task characteristics and test administrations can be standardized, but this may threaten the extrapolation by narrowing down the universe of generalization compared to the target domain; therefore it is important to keep a balance between generalization and extrapolation. Overall, generalizability increases as the sample size of observations for each test taker is increased and conditions of observations (i.e., test duration, setting, instructions, equipment, rating) are standardized. Reliability and generalizability studies require that test administrations and procedures are standardized on different occasions so that the variability resulting from different conditions of observations is reduced. If the test claim involves different groups of test takers (i.e., educational contexts, gender, L1), differential item functioning (DIF) analysis can be conducted.

Extrapolation relates to how much of the target domain the universe of generalization covers. The more similar the universe of generalization and the target

domain are, the more plausible the extrapolation becomes. The justification of extrapolation mainly relies on negative argument as in Popper's (1963) falsification principle. In other words, serious attempts are made to identify the differences between the target domain and the universe of generalization which would threaten the extrapolation inference. The evaluation of extrapolation is done through analytical and empirical evidence. Analytical evidence relates to the conceptual and theoretical relationship between the universe of generalization and the target domain and is mainly generated during the test development process. Evidence about students' perceptions of the relevance of the test to the proposed test interpretations and uses (i.e., face validity) can also be used in the analytical evaluation of extrapolation since a lack of such relevance can threaten the extrapolation inference (e.g., students not putting enough effort into the test because of the irrelevant appearance of the test). Empirical evidence examines the relationship between test scores and criterion scores from other measurements in the target domain. It involves correlational studies or regression studies if test scores are used to predict performance on the related criterion measure. Extrapolation is the weak link in objective tests since they don't usually involve high-fidelity simulation and authentic tasks (Kane et al, 1999). The tasks in universe of generalization should be as similar as possible to the tasks in the target domain without threating the required standardization of generalization.

The decision (or utilization) inference links the interpreted score to uses of the score (i.e., placement, selection) through a decision rule, and it is where test consequences are involved in validation. The analysis of the decision inference involves washback studies and stakeholder input about the test use and impact. Therefore, positive and negative consequences of a test should be examined mainly

through qualitative methods. Positive consequences are expected to outweigh negative consequences for justification of the decision inference.

2.4.4.3 Uses of Argument-based Validation in Language Assessment

Following Kane's argument-based approach, there have been studies to exemplify the types of evidence that might be needed for suggested test interpretations and uses (Bachman, 2005; Bachman, & Palmer, 2010; Chapelle et al, 2008; Kane, 2002, 2006; Drackert, 2016). Chapelle et al (2008) contributed to the understanding of validity argument. They showed how to apply an argument-based approach in validating TOEFL IBT for making decisions about test takers' readiness for academic language study at English-medium universities. They extended Kane's bridge analogy to six inferences by including domain definition and explanation, and they used Toulmin's argument model for each inference (see section 2.4.4.1). Later, Bachman and Palmer (2010) formed a new framework of Assessment Use Argument (AUA). AUA consisted of four claims of assessment records, interpretations, decisions, and consequences. Each of these claims was based on warrants which required backing and/or rebuttals. Following these developments, an argument-based approach was used to validate different types of language tests for placement purposes. For example, Li (2015) used an argument-based approach to validate the uses of an English Placement Test (EPT) by focusing on extrapolation and ramification (decision) inferences. Similarly, Gaillard (2015) validated a French EIT which can be used as an aural/oral component of a French placement test through an argument-based approach. Regarding language assessment in SLA, Drackert (2016) is the first example of showing how Kane's argument-based can be applied to the validation of a Russian EIT for uses in SLA.

These example studies show different possible validity arguments leading to test claims and uses depending on test context. They don't necessarily involve all chains of inferences in the argument-based approach. Rather, they focus on inferences which seem to be important for the particular test use. If an assumption underlying an inference is already plausible and evident, it doesn't need further evidence for the justification of interpretive argument. As Kane (1992, p. 530) said, "Validity evidence is most effective when it addresses the weakest parts of the interpretive argument". Argument-based approach helps to state weaknesses and threats underlying a test use by clearly defining score meanings. Overall, more practical studies are needed to exemplify the implication of argument-based approach for different test interpretations and uses.

2.4.5 Current Views in Validity

There is currently a widespread disagreement over whether the term validity should be narrowly defined and limited to test interpretations or should also involve test uses (for a detailed discussion see the special issue on Validity in Newton & Baird, 2016). If validity is used for test interpretations, validation should involve investigating the plausibility of test interpretations, and if validity is used for test uses, validation should involve investigating the appropriateness of test uses (Kane, 2016).

Nevertheless, there is more agreement over some aspects of validity and validation as summarised by Cizek (2016). First, validity relates to the interpretations or inferences based on test scores rather than tests themselves (Cronbach, 1971; Messick, 1989). Second, it advocates the use of a unitary concept of validity rather than diverse types of validity (Messick, 1989, 1995). Third, validity is seen along a continuum of evidences instead of a dichotomy (Zumbo, 2007). Fourth, validation is not a one-time activity, but an ongoing process (Shepard, 1993). Therefore, original

judgments may change during this process. Fifth, validation involves value judgements (Messick, 1975). Finally, validation of score interpretation is a necessary but insufficient precursor to justification of a test use (i.e., Cizek, 2016; Sireci, 2016). The current research has a "liberal" stand, in Newton and Baird's (2016) terms, arguing that validity can involve both test interpretation and / or test use depending on the test purpose. This stand is in line with the structure of argument-based approach because argument-based approach allows "to focus on an interpretation to the exclusion of any uses" (Kane, 2016, p. 207).

2.5 Validation in SLA Assessment and Educational Assessment

There are differences in the validation of assessment tools in SLA research and educational programs. Specifically, Norris and Ortega (2012) stated that validation studies in SLA should be linked to theories of L2 acquisition. They said that validation in SLA is different than it is in educational contexts because SLA assessment is focused on theory-based interpretations about L2 knowledge constructs rather than decisions and actions associated with educational assessment. On the other hand, language tests in educational programs such as placement tests should be correlated with the language curriculum, syllabus and course content (see, for example Norris, 2008 and Gaillard, 2015 for the validation of L2 placement tests).

Purpura, Brown and Schoonen (2015) argued that the amount and kind of evidence needed to justify a validity argument in SLA-research oriented assessment is dependent on the goal of the study, the research claims and consequences of using the scores for intended research goals. Regarding educational assessments, Norris (2008) suggested addressing the following questions:

> (a) who uses them, (b) what kinds of information they provide about whom or what, (c) why and how the information is sought, (d) what decisions and

actions are taken on their basis, and (e) what consequences are intended (and not intended) to occur as a result… (p. 73)

This proposed research aims to collect evidence for assessing the usefulness of the C-test to control general language proficiency of Turkish L2 learners in SLA research and predict test taker performance in TYS. The evidence will be both quantitative (i.e., test scores, self-assessment, L2 background and use) and qualitative (i.e., semi-structured interviews, stakeholder input) in accordance with test interpretations and uses. The methods will be explained in more detail in Chapter 4.

2.6 Summary

This chapter explained the gaps of measuring L2 proficiency through systematic and validated methods. As noted by several researchers, defining and measuring language proficiency has been undervalued in SLA context (i.e., Hulstijn, 2012; Norris & Ortega, 2003, 2012). Therefore, the current research first operationalises the language proficiency and then selects an appropriate measurement instrument. It conceptualises L2 proficiency as a unitary concept (Oller, 1979) involving the general components of L2 proficiency, namely grammar and lexis. This is because test results are reported as a single score indicating a short-cut estimate of language proficiency. The C-test is fit for this conceptualisation for several reasons. First, it can provide short-cut estimates of language proficiency since it involves the general elements of the language proficiency construct and tap into both receptive and productive skills to a certain extent. Second, it is free and can be completed within a short amount of time compared to commercial time-consuming proficiency tests. Third, it is relatively easy to develop, administer, and score, and thus, saves researchers from the time and energy costs. Fourth, it has provided high reliability and validity indices in various languages (i.e., Eckes & Grotjahn, 2006; Lee-Ellis, 2009; Norris, 2006, 2018; Son,

2018). Study 1 in Chapter 6 validates a Turkish C-test for SLA research purposes where researchers need to control the language proficiency of their participants with an independent proficiency measure.

Validating low-stakes screening and diagnostic tests has also been underestimated in educational context despite their impact when they are published online open to hundreds of thousands of learners (i.e., Alderson et al, 2015; Chapelle et al, 2003; Schmidgall et al, 2017). Therefore, study 2 in Chapter 7 validates a Turkish C-test as a quick screening test to inform students about their test readiness for the TYS.

The limitations of the C-tests that will be further investigated in the present research involve the ceiling effect and the low face validity. The ceiling effect, that is the insufficiency of the C-test to distinguish well between high-level proficient learners was observed in several studies (i.e., Grotjahn, 1987; Klein-Braley, 1985; Son, 2018); however, it's yet uncertain whether C-tests have enough discriminative power to distinguish among high-level proficient learners. This dissertation will further shed light into this aspect by recruiting Turkish L2 learners along a wide spectrum of proficiency. Regarding the face validity, the existing few studies on stakeholders' perception of the C-test showed that C-tests were not viewed positively by teachers and learners (Sigott, 2004). Study 1 and study 2 will further investigate stakeholders' (researcher, teachers, learners) perception of the C-test through the thematic analysis of interviews and open-ended survey answers. To the best of my knowledge, no prior study has investigated researchers' views of the C-test in a systematic way.

The adopted validation approach that is used to evaluate test uses is Kane's (2006) argument-based approach (see sections 6.2 and 7.2). Since argument-based

approach is pragmatic and starts the test validation process from test uses, it is the preferred method to validate the uses of the Turkish C-test as a research instrument to control language proficiency and a screening test to predict candidate readiness for TYS. It was preferred to other argument-based validation frameworks such as Chapelle et al (2008) developed for TOEFL IBT since it was more suitable and practical for the low-stakes purposes of the Turkish C-test.

CHAPTER 3: CONTEXT

3.1 Introduction

This chapter situates the assessment of L2 proficiency in the Turkish context by providing information about Turkish as an L2 in and outside of Turkey including Turkish SLA research and the development of the TYS to meet the demands of increasing number of Turkish L2 learners. By doing so, it justifies the need for the uses of the Turkish C-test as a research instrument and screening test for TYS (see section 2.3.2 for details about C-test uses). Furthermore, morphology (the study of word structure and word formation) of the Turkish language is discussed since it was important to consider in the development of the Turkish C-test. Since Turkish C-test is proposed to be used in academic settings by researchers and learners (as opposed to, for example, employment settings), this context chapter is limited to the uses of Turkish as an L2 in academic contexts.

3.2 Turkish as an L2

Although there are no official reports stating the total number of Turkish L2 learners in and outside of Turkey, there is a growing interest in teaching and learning Turkish as an L2 (Gürel, 2016). Within Turkish universities, the number of international students increased from 48,183 to 125,138 between 2013 and 2018 according to the 2018 statistics[8] reported by the Council of Higher Education of Turkey (YÖK). Among these international students, Syrian students comprise the largest group with 17%. This is followed by students from Azerbaijan (14%), Turkmenistan (10%), Iran (5%), Afghanistan (5%), Iraq (4%), and Germany (3%). Following concerns regarding the readiness of these international students to deal with the demands of academic instruction in Turkish, *Turkish Proficiency Exam* (TYS) was developed in

8 https://istatistik.yok.gov.tr/

2011 by Yunus Emre Institution in order to assess the Turkish proficiency of international students, who want to study at Turkish universities (details about TYS follow in section 3.2.2).

Outside of Turkey, Turkish L2 education is supported at the governmental level in several countries through language scholarship programs such as Critical Language Scholarship (CLS[9]) and The Language Flagship[10] (Gürel, 2016). The Bureau of Educational and Cultural Affairs in the US Department of State announced Turkish as one of the fourteen critical need foreign languages (FLs) and included it in the CLS program (Gürel, 2016). CLS provides students from the US an opportunity to learn Turkish through intensive language instruction both home and abroad. The US also has federal programs such as The Language Flagship, National Security Language Initiative for Youth (NSLI-Y), STARTALK, and National Council of Less Commonly Taught Languages (NCOLCTL) which promote learning critical need FLs. Due to this recognition and demand at the governmental level, the number of US higher education institutions offering Turkish courses was around 30 to 50 between 2009 and 2016 according to the reports of American Association of Teachers of Turkic Languages (Ergül, 2017, p.7).

The British Council has also recognized Turkish as one of the ten languages that are expected to be important in the UK for the next 20 years in Languages of Future report considering economic, political, cultural and educational factors (Tinsley & Board, 2013). Although, there is not an official report stating enrolment in Turkish courses or the number of universities offering Turkish courses in the UK, the researcher of this study was able to find and contact 9 universities offering Turkish

9 https://www.clscholarship.org
10 https://www.thelanguageflagship.org/

classes in the UK through an online search. These Turkish classes were mostly in UK universities offering Middle Eastern Studies courses or have a Modern Language Center.

Considering recognition of Turkish language and higher number of Turkish L2 courses offered in UK and US higher education institutions, these two contexts were chosen in the test development and validation study 1 where Turkish C-test is evaluated as a short-cut estimate of language proficiency in SLA studies. Balkan and European countries (i.e., Bosnia and Herzegovina, Germany, Greece, Romania, Netherlands) where Turkish is spoken at home as a minority or immigrant language are not involved since Turkish L2 learners who learn Turkish in academic settings were targeted. On the other hand, validation study 2 is conducted in all the countries where TYS is administered, since it evaluates Turkish C-test as a screening test for TYS and aims to reach a representative sample of TYS candidates. The following two subsections provide information about the language proficiency instruments currently used in Turkish SLA research and TYS as a standardized measure of Turkish language proficiency.

3.2.1 Language Proficiency Instruments in Turkish SLA Research

In Turkish SLA research, there is a lack of standardized and validated L2 proficiency tests that are freely available to SLA researchers. However, researchers of Turkish SLA need to have access to a standardized Turkish L2 proficiency test either to control for proficiency effects on L2 performance such as developing phonological competence and acquisition of causative verbs (e.g., Özçelik & Sprouse, 2016; Montrul, 2016) or to select participants with a certain proficiency level into their study (e.g., Bayyurt & Martı 2016).

In a recent volume of 11 experimental studies on SLA of Turkish (Gürel, 2016), only 4 studies assessed Turkish L2 proficiency independently with a test and the rest of the studies depended on institutional level, in-house testing, self-assessment, existing proficiency certificate, and the Language Experience and Proficiency Questionnaire (LEAP-Q). The tests used in these 4 studies included cloze test, multiple-choice cloze test, and a read-aloud task; however, none of these tests were standardized to enable generalizability, replicability, and interpretability across different studies. For example, Özçelik and Sprouse (2016) used a Turkish multiple-choice cloze test (Özçelik, 2011) to categorize Turkish L2 learners into three different proficiency levels when they investigated learners' dependence on auditory and orthographic stimuli in choosing the right suffix vowel across different proficiency levels. This multiple-choice cloze test was originally a Turkish translation of an English cloze test which was created based on a passage in an American advanced level course book (Montrul, 1997). Montrul (1997) stated that Turkish native speakers performed with 51.38% accuracy rate on this Turkish cloze test, which seems to be quite low, probably due to the fact that the text was originally English, and Turkish has a typologically different structure than English (see section 2.3 for a discussion on cloze-tests and section 3.3 for typological differences of Turkish).

Overall, the lack of a standardized and accessible Turkish L2 proficiency instrument is an emerging issue to generalize the results across studies in Turkish SLA (Gürel, 2016). The lack of uniformity in researchers' preference of a measure of Turkish L2 proficiency also contributes to the issues of generalizability, replicability and interpretability across different contexts. Validation Study 1 aims to address this issue by evaluating the Turkish C-test as an instrument to control for language proficiency in SLA studies in Chapter 6.

3.2.2 Turkish Proficiency Exam (TYS)

Regarding international students in Turkish-medium universities, in 2011, YÖK required universities to allow the registration of international students in their academic programs only if their Turkish proficiency is at B2 level or above according to the CEFR. Figure 5 below briefly summarises the CEFR scale.

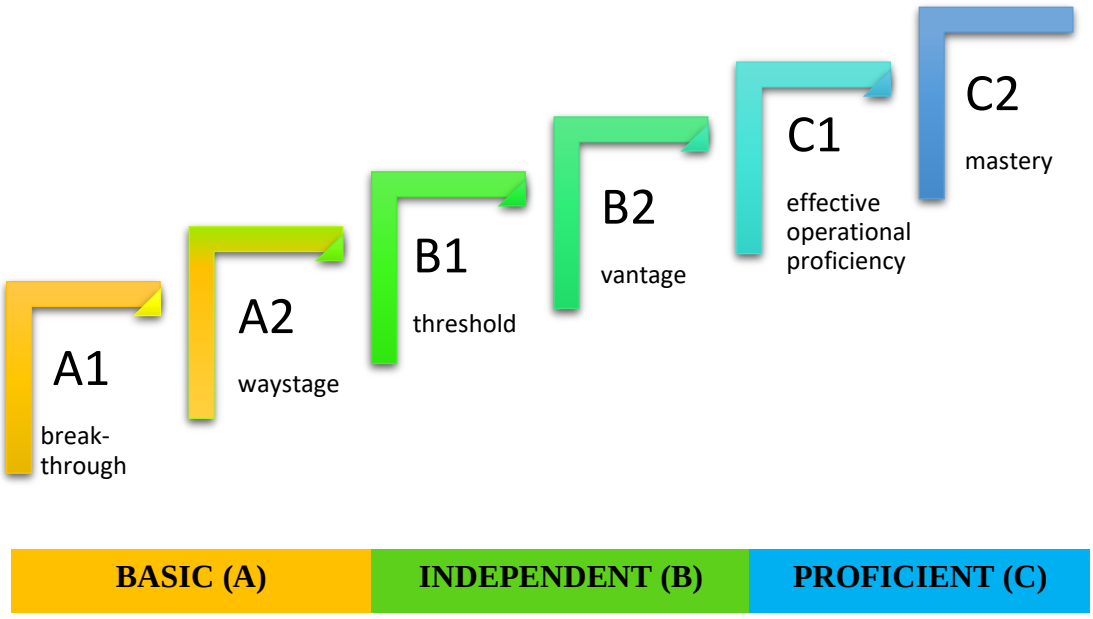


Figure 5. CEFR Scale and Levels

If international students were not at least at B2 level, they were required to take pre-sessional Turkish language classes prior to starting their academic degrees. However, there appeared to be no uniformity across universities in terms of the required Turkish proficiency for academic registration since each university conducted their own Turkish proficiency exams. For example, some universities required Turkish B2 level (on the condition of having C1 level at the end of the semester) and some Turkish C1 level as a prerequisite for university admissions. As a result, there were cases when a university did not recognize the language certificate given by another university. Therefore, in order to facilitate the enrolment and admission of international students into Turkish-medium universities through a

consistent and internationally recognized proficiency test in Turkish, the Yunus Emre Institution (YEE[11]) was authorized to develop and administer the *Turkish Proficiency Exam* (TYS) in 2011. YEE is a non-profit cultural institution which has headquarters in Turkey and 59 exam centres in a total of 48 different countries. The first worldwide administration of the TYS was conducted in 10 countries (Kosovo, Albania, Egypt, Iran, Azerbaijan, Japan, Georgia, Bosnia-Herzegovina, Belgium, Kazakhstan) between 24 and 25 May 2013. A total of 10,989 candidates have taken the TYS as of January 2019. Although the overall percentage of successful candidates has not been publicly shared, it was announced as 65% for the TYS that took place in January 2018 and 88% for the one in January 2019[12]. The highest number of candidates was from the following five countries in order: Azerbaijan, Turkey, Kazakhstan, Iran, and Kosovo.

The TYS price depends on the local wage rate and the local currency of the country where the exam is conducted (i.e., 50 euros for candidates in Bosnia-Herzegovina, 100 euros for the ones in Germany). The exam price does not seem to be a very big amount, but if the candidates are not at least at B2 level and have to take the TYS a few times, they may not want to pay the exam fee several times. Furthermore, they can only take the TYS three times a year (January, May, or between July and September considering the dates of university admissions) in specific locations, which adds a degree of cost and inconvenience to taking TYS several times.

[11] YEE stands for Yunus Emre Enstitüsü
[12] https://yee.org.tr/tr/haber/2018-tys-sonuclari-aciklandi
https://www.yee.org.tr/tr/haber/tysde-basari-orani-88

Considering the costs of taking TYS and the need to pass it to study at Turkish universities, students and educators are in urgent need of a low-cost screening test that would help to determine whether a student is ready to take TYS (i.e., whether they are at least B2 level). This would help to reduce the time, money, and energy costs. Therefore, validation study 2 evaluates the Turkish C-test, which can be taken online, for free, and at any time, as a screening test for TYS. In order to establish the relationship between these the C-test and TYS, the following subsection gives background information about TYS.

3.2.2.1 TYS Purpose and Description

The main purpose of TYS is to assess the language proficiency of Turkish L2 speakers (i.e., international students in Turkey) in four language skills (reading, writing, listening, speaking) for their registration at Turkish universities or employment at governmental level (i.e., Ministry of Health) in Turkey. In addition to these uses for Turkish L2 speakers, TYS can also be used to assess the language proficiency of Turkish L1 speakers if they want to work outside of Turkey as a teacher or translator of Turkish. Given that TYS is a language proficiency test, C-test as a short-cut estimate of language proficiency would be useful to screen candidates for TYS (see section 2.3.2.2 for details about what C-tests measure).

TYS was developed by following a TYS Framework Program, which was produced by adapting CEFR to Turkish language. It is a three-hour paper-based examination consisting of three stages: the first stage involves a 60-minute reading section and a 45-minute listening section; then, after a 15-minute break the second stage involves two writing tasks lasting a total of 60 minutes; the third stage, which is conducted after a one-hour break or on the following day or days depending on the number of candidates, involves two speaking tasks lasting a total of 15 minutes.

The reading section involves six different texts with 40 relevant questions, and the listening section involves six texts with 30 relevant questions. Listening and reading texts have different genres ranging from academic texts to advertisement texts, which was also the case for the C-test texts (see Chapter 5 for details about C-test texts). They aim at measuring test takers' ability to understand the given texts in written and oral modes and respond to relevant questions. Questions include matching task, fill-in the gaps task, true/false questions and multiple-choice questions. The writing section involves two tasks. The first one is a guided writing task such as writing an answer to an e-mail and writing a report based on given graphics, and the test takers are asked to write a text with a minimum of 125 words as a response to the given prompts within 20 minutes. The second is an argument essay task based on a given topic, and the test takers are asked to write an essay with a minimum of 200 words within 40 minutes. The speaking section also comprises two tasks. For the dialogue task, the test taker chooses a random speaking card and answers 7 questions directed by the examiner on the topic written on the card for 10 minutes. For the long-turn speaking task, the test taker is asked one question about the topic on the speaking card. The test taker is given 2 to 3 minutes to prepare his/her talk, and then s/he is expected to talk about the given topic independently for 5 minutes.

The TYS is evaluated over 100 points with each section contributing 25 points to the total score. If the candidates are successful, that is they achieve a minimum score of 55 in total with at least a score of 12.5 in each section, they are given the Certificate of Turkish Proficiency, which is valid for 2 years. The Certificate of Turkish Proficiency has three different levels according to the CEFR, which is shown on the Figure 6 below with the corresponding TYS scores.

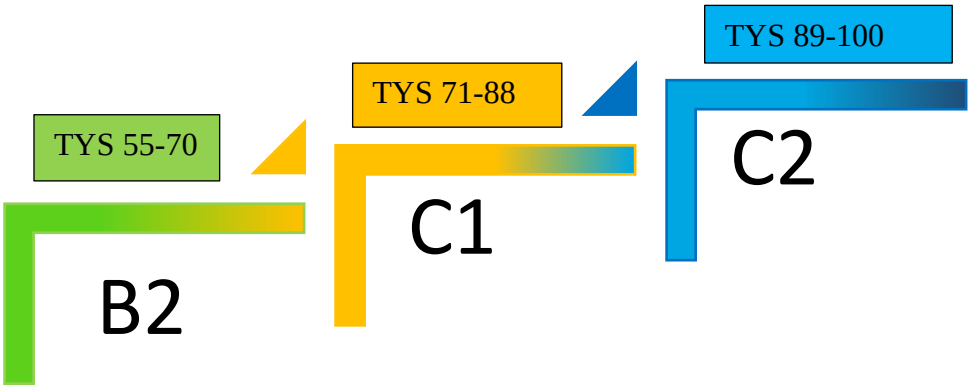


Figure 6. CEFR levels and TYS scores

As seen in Figure 6, B2 is for 55-70 points, C1 for 71-88 points, and C2 for 89-100 points. Below B2 level, no attempt is made to differentiate students, and no certificate is given. These cut scores are used for conducting regression analyses between C-test and TYS in Chapter 7.

3.3 Morphology of the Turkish Language

Turkish belongs to the Turkic language family; thus, it is typologically different than Indo-European languages in which C-test research has been extensively conducted (see Grotjahn, 2017). These typological differences include extensive agglutination (forming complex words by stringing meaningful linguistic units together) and morphological productivity (coining new words by using existing linguistic units). Besides these, Turkish is a pro-drop (allowing dropping some pronouns) and free word order (flexible word order in a sentence) language which prevents straightforward C-test deletion. Therefore, this research is unique in showing how the challenges resulting from these morphological features of Turkish are addressed in developing a Turkish C-test (see Chapter 5).

One of the most prominent features of the Turkish language is its highly agglutinative structure. Since it is quite different than some most spoken languages such as English, Lewis (2000, p. xx foreword) stated the extensive agglutination of

Turkish as what “English-speakers find most alien”. Agglutination means deriving complex words by adding morphemes to a word in a way that segmentation (or separation) of morphemes and the word is relatively easy after their union. Morphemes are the smallest meaningful linguistic units that cannot be further divided such as the plural marker *-s* and the word root *dog* in the word *dogs*. To explain agglutination in a Turkish example[13], the Turkish word *evlerinizden* (from your houses) is derived by adding three morphemes (plural marker *-ler* (-s), possessive marker *-iniz* (your), ablative marker *-den* (from) to the word root *ev* (house) as exemplified below:

Ev	+	*ler*	+	*iniz*	+	*den*
House		plural marker		possessive m.		ablative m.

As a result of its highly agglutinative structure, Turkish is morphologically very productive, which means that new words can be coined by using the already existing words and morphemes. Although some Indo-European languages such as English also allow morphological formation such as prefixation and suffixation (i.e., un-happy, dis-courage, care-less-ness), their productivity is still limited compared to highly productive morphology of Turkish. For instance, Turkish multimorphemic words (having more than one morpheme) are estimated to have an average of 4.8 morphemes (Hankamer, 1989) while English multimorphemic words are estimated to have an average of 2.5 morphemes (Balota et al, 2007). In Turkish, it is possible to form a sentence with one word as in *sakinleştirebildim* which translates to English as *I was able to calm myself/you/her/him/it/them down* in the form of a 7-word sentence. The segregation of this sentence (word) is stated below.

Sakin + *leş* + *tir* + *ebil* + *miş* + *im* + *dir*
Calm + derivation + causative + modality + past perfect + person + modality

[13] Turkish words are stated in italics, and their English meanings are given in parenthesis.

The sentence above also shows the pro-drop feature of Turkish by omitting a subject and an object pronoun. While the subject pronoun *ben* (I) can be easily inferred from the personal marker *–(i)m*, the object pronoun is not obvious, and it could be *kendimi* (myself), *onu* (her/him/it), *seni* (you), or *onlari* (them). The object pronoun can only be inferred from the pragmatic clues in the context (discourse). Hence, Turkish is a pro-drop language which allows subject and object pronouns to be omitted when they can be inferred from verbal inflections or pragmatic clues.

Although the common word order in Turkish sentences is subject-object-verb (SOV), Turkish is also a free word order language where constituents of a sentence can scramble freely depending on which constituent is emphasized. Free word order of Turkish sometimes leads to inverted sentences as seen in the example below.

Ankarada'dan	abim	gelmiş. (inverted)
From Ankara	my brother	came.

Abim	Ankarada'dan	gelmiş. (usual SOV order)
My brother	from Ankara	came.

While the second sentence shows the usual SOV order, the first sentence shows the inverted version, which is completely acceptable if the subject *abim* (my brother) is the emphasized constituent of the sentence.

3.3.1 Challenges in Turkish C-test Research

There is, as yet, little research investigating whether the C-test principle is applicable to Turkish given the typological features of Turkish explained in section 3.3 (see Grotjahn, 2017). Therefore, showing how the C-test principle is applied to Turkish step by step with several language-specific factors is one of the contributions of this dissertation (see Chapter 5 for test development).

Daller et al. (2002) highlighted certain challenges that arise in the application of the classical C-test principle to the Turkish language (see section 2.3.2 for details about C-tests). These challenges are summarised below.

First, as the grammatical information is generally encoded in suffixes (morphemes added to words) in Turkish, most of the grammatical information is deleted with the second half deletion principle. A study conducted by Baur and Meder (1994) with two native Turkish speakers found that morphological structures deleted using the second-half deletion method often cannot be restored. For this reason, Grotjahn (1987) suggested that it might be reasonable to delete the first half of a word, or the middle of a word for languages typologically different than English, such as Turkish or Hebrew. However, Cleary (1988) found that the deletion of the first half discourages learners from utilizing the text. Furthermore, for the languages that are processed online from left to right, the second-half deletion should also be more psycholinguistically (psychological factors affecting learning a language) valid than first-half deletion (Lee-Ellis, 2009). Also, deleting the first half would cause more content word (words with clear meanings compared to grammatical/function words) deletion since Turkish is agglutinative and the test can turn out to be more difficult than expected.

Second, Turkish texts contain relatively little redundant information compared to languages like English and German. Therefore, the reconstruction of C-test items may be more difficult, even for native speakers. Note the example of one-word sentence *sakinleştirebildim* and its seven-word equivalent in English *I was able to calm myself/you/her/him/it/them down* in section 3.3

Another issue with Turkish C-tests is that there are fewer independent function/grammatical words than content words in Turkish texts, as the grammatical

information is typically provided in suffixes added to content words (Daller *et al.*, 2002). Note the example *ev-ler-iniz-den* (from your houses) in section 3.3 where grammatical information is provided in the form of three suffixes *(-ler, -iniz, -den*) added to the content word *ev* (house) while it is in the form of two independent function words (from and your) and one suffix (-s) in English. As content words are more difficult to restore, a Turkish C-test might prove more difficult in this respect as well.

For these reasons, Daller et al. (2002) claimed that a Turkish C-test constructed with second half deletion method might be more difficult even for native speakers. To overcome this challenge, researchers suggested alternative deletion principles to investigate whether they are more applicable to Turkish. In the next section, these deletion principles and their results with Turkish L2 learners will be summarised.

3.3.2 Deletion Principles in Turkish C-test

In the first example, Baur and Meder (1994) compared the second-half deletion principle with the syllable principle, in which every third syllable is deleted. They found that, among Turkish-German bilingual children, the texts deleted using the second-half deletion method were equally difficult to restore as texts deleted with the syllable deletion method.

Subsequently, Daller *et al.* (2002) used the classical second-half deletion principle alongside alternative principles, including the morpheme principle, syllable principle (which was explained above), middle principle, and third half principle in order to compare their levels of difficulty. The similarities of these approaches are that all start from the second sentence onwards, and first and last sentences are left intact. The differences are described as follows: The morpheme principle is the

deletion of every third morpheme; therefore, units of deletion are morphemes rather than words as in the second half principle. The middle principle is the deletion of the middle of every second word instead of first-half or second-half deletion. Thus, half of the total number of letters is deleted in the middle. If the total number of letters is an odd number, one is subtracted from the odd number and then it is divided by two. However, the authors don't mention how they coped with two-word letters such as the word *ve (and)*. Finally, the third half principle is the deletion of the second half of every third word rather than every second word. These differences are exemplified by using different deletion principles on the same sentence below:

Okulumuzda	yabancı	dil	eğitimine	büyük	önem	verilir (undeleted)
In our school	foreign	language	education	great	importance	is given.

Oku + l + umuz + da	yaban + cı	dil	öğret + im + i + ne	büyü + k
read- noun- poss- loc	foreign- adj	language	teach-noun-poss-dat	greaten -adj

önem	ver + il + ir. (segregation of words)
importance	give-pass-present

Okulumuzda yab___dil öğret___ büyük ön___ verilir. (second-half principle)

Okul___da yaban___dil öğret___ine ___k önem ___ilir. (morpheme principle)

Oku___muzda ___bancı ___ öğre___mine ___yük ö___ veri___. (syllable principle)

Okulumuzda ya___cı dil öğr___ne büyük ö___m verilir. (middle principle)

Okulumuzda yabancı d__ öğretimine büyük ön__ verilir. (third half principle)

As seen in the examples, there are discernible differences between the deletion methods. In the morpheme and syllable principles, roots of the words are sometimes deleted in comparison to the second half and third half principle, which might lead to the proliferation of alternative solutions and make the test more challenging. For example, instead of *büyük (big)*, *çok (much)* is possible and *gösterilir (shown)* is an alternative to *verilir (given)* in the morpheme principle when the word roots are

deleted. The middle principle might help to solve this problem since part of the roots and suffixes are deleted. Another issue regarding morpheme and syllable principle is that the word is sometimes deleted completely when it consists of one morpheme (i.e., önem, için, sebep) or one syllable (i.e., dil, zor), which might make C-Test solving more challenging. The third half principle seems to be the least challenging one since it has the least deletions. However, it is questionable how much deviation from the C-Test construct there is due to changing the deletion ratio.

Daller *et al* (2002) found that the morpheme deletion principle should be excluded, as even adult native speakers were unable to complete texts deleted using this principle with scores lower than 75% accuracy. They also found that syllable and middle deletion principles are useful alternatives, and that the second-half deletion principle is applicable. However, while their tests showed good reliability with Turkish-German bilingual children, the validity of the tests remained in doubt, as highly proficient monolingual adult native Turkish speakers could not restore 95% of the deleted words accurately for all texts with different methods. Furthermore, the results obtained for bilinguals may change if the C-test were administered to learners of Turkish as a foreign language in academic settings.

More recently, Caprez and Gönç (2006) developed the explorative method and the first-suffix method as alternatives to the second-half deletion method considering the agglutinative structure of the Turkish language. They aimed to minimize the number of possible acceptable solutions to the deleted items. According to the first suffix principle, the first suffix after the root word is deleted in words with multiple suffixes; the last two syllables are deleted in monosyllabic words; and, the second syllable is deleted in two and polysyllabic words. In addition, depending on the difficulty level of words, the second syllable, minus one, is deleted in the words ***değil***

(not), *eğer* (if), and *için* (for). Regarding the explorative method, the authors used a mixture of syllable and morpheme deletions. They found that native Turkish speakers were able to complete 95% of the deleted words correctly, and a satisfactory reliability index was reached with both methods. The average mean was slightly higher for the C-test developed with the explorative method, although this may be attributed to the educational background of the group who took this test, compared with the other group, who took the test using the first-suffix principle. Regarding the bilingual participants, both Turkish bilingual children from Switzerland and native Turkish children from Turkey performed better in the C-test developed via the first suffix principle (M=48.9, 71.8) compared to C-Test with explorative method (M=43.5, 70.2). This might be because word roots are not deleted or only partially deleted in the first suffix principle in comparison to explorative method where word roots are sometimes deleted. As a result, the researchers found the new first suffix principle to be an effective method for developing C-tests.

In the final example, Sağın-Şimşek (2006) developed and used a two-text Turkish C-test using the second-half deletion method, together with German and English C-tests, to investigate the role of Turkish as an L1 and German as an L2 in the acquisition of English as an L3 in a German school. According to the results, Turkish-German bilinguals performed the highest scores in Turkish and the least scores in German although they had been learning German for longer than English. Sağın-Şimşek attributed this result to the different acquisition and learning conditions of languages; while students acquired German through oral communication in daily life settings, they learned English through instruction, in an academic setting. This is important evidence that the way a language is learned, whether in an academic setting or in daily life, makes a difference to learners' performance in C-tests. Based on this

evidence, only the learners who learned Turkish in classroom settings were recruited in test development and validation study 1 since C-test is in written format and require certain literacy skills.

Based on previous research, it seems that second-half deletion was found to work equally difficult as alternative deletion principles when the Turkish C-tests were administered to bilingual pupils. This provides a basis for the assumption that second-half deletion method has the potential to be administered to adult Turkish second/foreign language learners despite its challenges, and it is the chosen deletion method in this dissertation to abide by the C-test rules as much as possible and, thus, not to deviate from the C-test construct.

3.4 Summary and Motivation for the New Turkish C-test

Overall, there has recently been a growing interest in Turkish L2 education in and outside of Turkey due to global developments such as scholarship and exchange programs (Gürel, 2016). Therefore, improvements are needed in the field of Turkish as a foreign language in order to meet the demands of increasing number of Turkish language learners. This dissertation will address several gaps in Turkish C-test research and more broadly address the issues regarding proficiency assessment in Turkish SLA research as well as educational assessment in Turkish.

First, the existing few studies on Turkish C-test focused on testing Turkish-German bilingual pupils and did not involve Turkish adult learners who learned Turkish through academic instruction. This is not a surprising finding considering C-test originated in German language and Germany, where there is a large population of Turkish-German bilingual children. However, as Sağın-Şimşek (2006) showed, the way a language is learned, whether in a classroom or in daily life from parents, makes a difference to learners' performance in C-tests. Therefore, this study makes a new

contribution by involving adult Turkish L2 learners, a population very different from the school-aged bilingual children and commonly recruited in Turkish SLA research (i.e., Gürel, 2016; Montrul, 1997; Özçelik, 2011). Second, in previous research, an acceptable accuracy rate with native adult speakers before administering the test to bilinguals was not always sought. There was one study (Daller, et al, 2002) in which native speakers only achieved scores of less than 75% accuracy, which does not seem to be close to the acceptable 90% accuracy level of native speaker accuracy suggested by Klein-Braley (1985). This dissertation ensures that C-test texts are piloted with native speakers first, and the texts with at least 80% accuracy level are selected to be administered to language learners. The reason why 80% threshold was chosen over the suggested 90% threshold is due to the initial findings in the test development stage (see section 5.7 for a discussion). One more point lacking in previous research is that it wasn't explored whether a Turkish C-test can be successful in distinguishing learners with different proficiency levels, and rather the focus was on the applicability of alternative deletion methods to the Turkish language without going through the test development and administration stages (i.e., Baur & Meder, 1994; Daller et al, 2002). Nevertheless, they provided support for the suitability of the classical second-half deletion method in Turkish.

This dissertation therefore uses second half deletion method without comparing it with alternative methods. It involves texts with different levels and investigates whether the Turkish C-test can distinguish learners with different proficiency levels. Furthermore, this study is the first to explain the development stages of a Turkish C-test (i.e., text selection, input from experts regarding texts, deletion strategies). It shows how to develop a Turkish C-test step by step, from text selection to word deletion with language-specific factors. It explains how to address

the unique challenges posed by the typological structure of Turkish when words are deleted. Therefore, it not only provides a free Turkish C-test open to public use, but also serves as a guide for future researchers or teachers who would like to develop their own C-tests in Turkish for different uses (see Chapter 5 for test development).

In a broader context, the Turkish C-test addresses the issue regarding the lack of a standardized Turkish proficiency test that is freely available and accessible to researchers, which makes it difficult to generalize results across studies (Gürel, 2016). Validation study 1 of this dissertation (Chapter 6) attempts to validate the use of the Turkish C-test as an instrument to control language proficiency in SLA studies. In doing so, the study shows that the Turkish C-test can potentially be used to generalize results across research studies.

This dissertation also addresses a gap in educational assessment in Turkish by validating the Turkish C-test as a low-cost screening test for TYS in validation study 2 described in Chapter 7 . It is the first study to evaluate the predictive power of a Turkish C-test to estimate levels set by a standardized proficiency test (TYS) and thus investigate its use as a screening test (see section 2.2.3 for the uses of screening tests). Since TYS is costly to take a few times, if candidates are not successful, considering that it is expensive and can only be taken in certain test locations at specific times, candidates can save time, money, and energy by using the online and free Turkish C-test in order to check their exam readiness for TYS (i.e., whether they are at least at B2 level to pass TYS). Hence, Turkish C-test will be useful to help candidates to decide whether TYS is appropriate for their level.

CHAPTER 4: METHODOLOGY

4.1 Introduction

This chapter presents the methodological choices used across the two validation studies as well as test development in order to answer the evaluation questions and justify the choice of those methods. Firstly, in order to justify the choice of the methods, the underlying epistemology and the philosophical assumptions of the validation studies are described. Then, data collection procedures common to both validation studies and background information about the data analysis methods (Item Response Theory and Thematic Analysis) follows. Full details about participants, instruments, data collection and analysis are provided in Chapter 5 for test development, Chapter 6 for validation study 1 and Chapter 7 for validation study 2.

4.2 Philosophical Background

Researchers should consult theoretical perspective and epistemology to justify their choice of methods and methodology; however, most studies do not state their philosophical assumptions clearly (Fulcher, 2014). Fulcher (2014) categorized epistemology (knowledge of how we know) in language testing under two groups as realist and anti-realist. Then, he distinguished two kinds of anti-realist stances: constructionism and instrumentalism.

Realism claims that what is observed and tested exists independently of the human mind. Therefore, it ignores the effect of test taker characteristics and test designers as well as stakeholders. It requires validity questions to investigate whether the construct (i.e., fluency, accuracy, complexity) in question in fact exists in the real world, and whether the differences in observed scores are linked to differences in the construct. It also requires strong testable theories, which are acknowledged not to be available in language testing or psychology (Fulcher, 2014).

Regarding anti-realist stances, constructionism claims that constructs do not

exist in the real world, rather they are socially constructed and contingent on ideologies. It emphasizes social impact and policy roles of tests. Fulcher interprets constructionism as pessimistic since it considers tests as mechanisms of power exercise and regards everything as an evidence of challenge. Instrumentalism, on the other hand, takes a middle position between realism and constructionism, and it does not make assumptions about the existence and necessity of constructs for language tests. Rather, it evaluates tests in terms of their practical consequences and usefulness. The current validation studies are epistemologically instrumentalist, in that, they take test uses and interpretations as a starting point of validation. They include test impact through stakeholder (test taker, instructor, and researcher) judgments into the validation process of the uses of the Turkish C-test as a research instrument and screening test. Kane's argument-based approach to validation is suitable for this purpose since its focus of validation is an interpretive argument which links observed test performance to test use and interpretation through a bridge of inferences (see section 2.6.4 for details about argument-based approach).

The current validation studies depend on the observable event of test taking and take a certain level of objectivity (Crotty, 1998). While the choice of the C-test as a measurement instrument is dependent on the researcher's observation that the research on this area is needed, the extent to which C-test can be used to predict and distinguish different levels of L2 learners in Turkish language context will be investigated. It is assumed that the difference between proficient and less proficient learners will be reflected in their test scores. Participants' C-test scores are supposed to support their proficiency levels according to their scores on the other measures of language proficiency (i.e., TYS, institutional level). Nevertheless, information on other factors that may contribute to test scores such as individual test taker

characteristics will also be collected through a background questionnaire feedback survey. Furthermore, stakeholders' attitude towards the C-test and how this affects their willingness to use the test will be revealed through semi-structured interviews and feedback surveys.

4.3 Overall Research Design

The current validation studies adopt a mixed-methods approach aiming to benefit from complementary strengths of both quantitative and qualitative data to validate the uses of the Turkish C-test under an argument-based approach adopted in this study. Table 5 shows the characteristics of samples and data sources used across three empirical chapters.

Table 5. Samples and data sources across test development and validation studies

	Test development	Study 1	Study 2
Samples	• 19 Turkish L1 speakers • 37 Turkish L2 learners	• 85 Turkish L2 learners • 10 SLA researchers (N=5 follow up interview)	• 79 TYS candidates (N=13 follow-up interview) • 34 instructors of Turkish (N=2 follow-up interview)
Data sources	• [14]11-text Turkish C-test • Background questionnaire • Post-test questionnaire	• 6-text Turkish C-test • Background questionnaires for learners and researchers • Surveys for learners and researchers • Interview questions for researchers	• 8-text Turkish C-test • Background questionnaires for candidates and instructors • Surveys for candidates and instructors • Interview questions for candidates and instructors

[14] The number of texts in each study are different based on the test purpose and analyses conducted. See sections 5.4, 6.4.2, and 7.4.2 for each sample of test versions

The target population of test takers is Turkish adult L2 learners with a wide range of proficiency level in test development and Study 1, and they are required to have learned Turkish in instructional settings in USA or UK. Since test development and study 1 investigates whether Turkish C-test can distinguish among learners with different proficiency levels, participants are sampled from university Turkish L2 classrooms ranging from beginner to advanced levels. On the other hand, the target population of test takers is Turkish L2 learners who have taken TYS worldwide in Study 2. Although the main purpose of the study 2 is to identify learners below or above the B2 threshold, its sub-aim is to predict all TYS levels based on C-test scores. Therefore, participants are sampled from all levels of TYS candidates.

Figure 7 below shows the overall research design involving inferences of the argument-based approach and the relevant data sources as well as data analysis for each inference. Note that the test development is a part of the overall validation process and its stages (i.e., text selection, word deletion) are involved under the *scoring* inference.

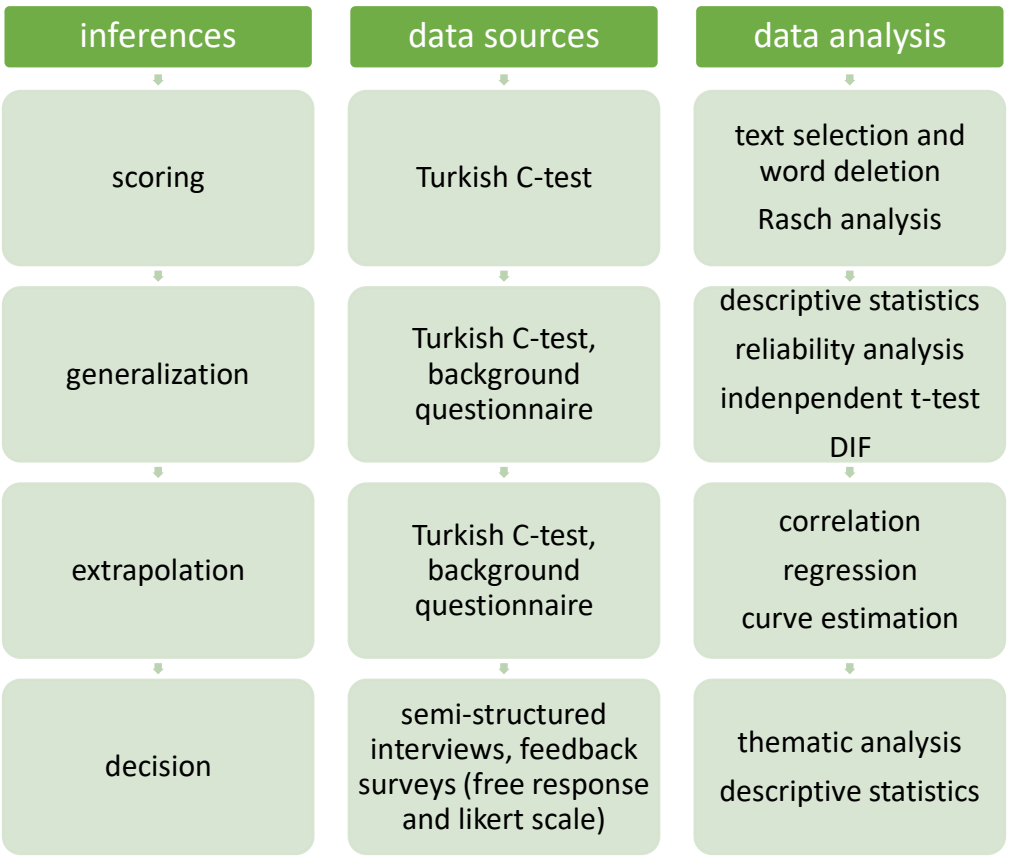


Figure 7. Overall Research Design

As seen in Figure 7, quantitative data sources comprise a large amount of all the data. On the other hand, qualitative data is relatively small; however, they still help to interpret and inform the quantitative data by involving stakeholders' perceptions into the validation process of test uses. This is particularly useful for evaluating the *decision* inference of the argument-based approach since the impact of the test on stakeholders can be involved in the decision through stakeholder input.

Quantitative data, involving test scores, participant background information, and feedback survey, provide the following information: (1) statistical information about psychometric characteristics of texts relating to *scoring* inference, (2) generalizability of test scores relating to *generalization* inference, (3) relation of test scores with other proficiency indicators relating to *extrapolation* inference, (4) statistical information about stakeholders' Likert-scale responses about their test

perception relating to *decision* inference (see section 2.6.4.1 for the explanation of these inferences). On the other hand, qualitative data, involving free response questions of the feedback survey and semi-structured interviews, bring a richer interpretation to test usefulness from stakeholders' perspective regarding the *decision* inference. For example, while the quantitative analysis of C-test and TYS scores will provide details about how accurately the Turkish C-test predicted TYS levels for extrapolation inference of validation study 2, qualitative analysis of interviews and free response survey questions will inform whether or why TYS candidates would use the Turkish C-test prior to TYS to predict their TYS levels and see their exam readiness for TYS for the decision inference. Eventually, if the test discourages test candidates or creates frustration among them, they will not use it. Furthermore, candidates' perception of the test might influence their performance in a way that if the test discourages them, they might finish it quickly and not do their best.

The adopted mixed method design of the validation studies is sequential explanatory which is characterized by quantitative data collection and analysis followed by qualitative data collection and analysis (Creswell & Zhou, 2016). Semi-structured interviews will be conducted after the collection and analysis of test scores, background questionnaires, and feedback surveys. By doing so, quantitative data will guide the interview questions. For example, if a test taker did very well on one of the most difficult texts despite having an overall low score, s/he will be asked questions about that text such as whether s/he is familiar with the topic of the text.

4.4. Data Collection

Initial investigation of the test development (piloting) will be conducted in paper and pencil format in learners' usual classroom time (see Chapter 5 for details and reflection on how it guided the validation studies). Following test development, both

validation studies will be conducted online without supervision to be able to reach a larger sample size. This type of test administration through the internet without a human proctor is called unproctored internet testing (UIT). UIT is commonly used in employment settings and distance-education institutions (i.e., Do, 2009; Makransky, & Glas, 2011; Rios, Liu, 2017).

Online data collection has several advantages. First, it allows to reach a larger number of participants considering the status Turkish as a LCTL. Second, it reduces the need to set up necessary sources (i.e., arranging a time and lab environment) for test administration, and thus, it gives test takers an opportunity to take the test anytime anywhere as long as they had an internet connection. Third, it makes reaching a very heterogeneous sample in terms of L2 background variables possible (see sections 6.3 and 7.3 for participant characteristics) by involving learners from a large number of institutions in different countries. Fourth, it also allows learners to see where they made mistakes and get their scores at the end, thus, learners have an opportunity to learn from their mistakes. Finally, it allows complete volunteering in participation without any pressure compared to methods such as conducting the study in L2 learners' usual class times. Therefore, it is expected that only interested people will take the test. It is acknowledged that there are also several limitations that come with online test administration which will be discussed in the final general discussion and conclusion chapter (see section 8.4).

4.5 Data Analysis Methods

Test scores will be analysed by using two measurement models which are Classical Test Theory (CTT) and Item Response Theory (IRT). CTT analysis will be used when the overall internal reliability of the test and standard error of measurement is calculated, and IRT analysis will be preferred to estimate item characteristics.

4.5.1 Classical Test Theory

CTT is a traditional approach to analysing test data. According to CTT, the observed test score for a person is obtained when random (unsystematic) measurement error is added to the person's true score. The reliability of using the observed test scores as estimates of the unknown true scores of interest is then defined as the proportion of true score variance to observed score variance. The reliability of a test can therefore range from 0 to 1 with scores above .80 commonly deemed acceptable. Since one's true score without measurement error cannot be obtained, CTT provides the following ways to calculate the reliability: 1) test-retest reliability, 2) internal consistency reliability. Test-retest reliability is calculated by administering the same test or parallel forms of the same test twice and correlating test takers' scores on both occasions. On the other hand, internal consistency reliability is calculated by administering a test only once and correlating test taker's scores on different items of the test. Cronbach's alpha is the most common measure of internal consistency reliability, and it is used as an estimate of the overall reliability in this study when each C-test text is treated as a polytomous item to be scored out of 20. However, Cronbach's alpha assumes that each item (text in this context) has the same difficulty level, and this might problematic in this study since C-test involves texts with different difficulty levels. Therefore, IRT reliability is also calculated, which will be described in the following section 4.5.2.

CTT also enables us to estimate item facility or item difficulty (proportion of test takers answering the item correctly) and item discrimination (to extent to which the item discriminates high scorers from low scorers) for dichotomous items. CTT will not be used to estimate item facility and discrimination in this study since C-test scores are polytomous and CTT only gives "sample-based descriptive statistics"

(Bachman, 2004, p. 139), which means item facility and discrimination values are based on a particular sample of test takers and items. Thus, making generalizations to other samples or parallel item formats may be difficult. IRT is proposed as an alternative to this limitation of CTT. In this study, item facility (difficulty) and discrimination values are calculated by using IRT since it estimates item parameters independent of the sample of test takers as well as test taker ability estimates independent of the particular sample of items. This means that although the data is taken from a particular sample of test takers, it is generalizable to other samples within the population. CTT is only used to estimate the overall reliability of the test and the standard error of measurement.

4.5.2 Item Response Theory

IRT is a more generalizable approach commonly preferred for large-scale assessments (Ellis & Ross, 2014). IRT models links test takers' performance on a specific item to their ability level and provides item quality statistics by putting both item difficulty and test taker ability estimates on the same scale. IRT models therefore model the probability that a test taker will achieve a particular score on a given item of a certain difficulty based on his/her ability and other item characteristics as represented by various item parameters. The common estimation scale enables us to map item parameters and test taker ability graphically and compare their distributions against each other (Baghaei & Grotjahn, 2014). Item parameters are "statistical estimates of a population based on the performance of a sample of test takers" (Bachman, 2004, p. 141). They can involve item difficulty, item discrimination, and guessing.

IRT analysis is commonly used in C-tests which typically include 4 to 6 texts (see for example, Baghaei, 2008a, 2008b; Baghaei, Grotjahn, 2014; Eckes, Baghaei, 2015; Lee-Ellis, 2009; Norris, 2006). A crucial assumption of the IRT analysis is the

conditional (local) independence assumption which means that conditional on test taker ability the test takers' responses to the items in a test should not be related to each other. However, on C-test texts, items (gaps) are semantically related to each other, in other words, whether you fill the 5th gap correctly may depend on whether you fill the 4th gap correctly. Therefore, while using IRT in C-test research, it is recommended to enter aggregated scores rather than individual gap scores as the unit of analysis in order to avoid the positive dependence of the individual items (Eckes & Baghaei, 2015). Therefore, in this dissertation, each C-test text is considered as a 20-point *superitem* (i.e., Grotjahn, 1987; Norris, 2006; Raatz, 1985) since each text has 20 gaps to be filled in and, for this reason, is scored out of 20.

4.5.2.1 IRT Models

IRT models have different types depending on a number of factors such as the number of underlying dimensions or traits that are being measured and the number of item parameters. Most traditional IRT models assume unidimensionality, which means that all items measure the same underlying trait (i.e., general language proficiency), while multidimensional IRT models also exist (Bachman, 2004). Most common IRT models are divided into the following three categories depending on the number of item parameters: 1-parameter IRT model (known as Rasch model) involving item difficulty parameters; 2-paramater IRT model involving item difficulty and discrimination parameters; and 3-parameter IRT model involving item difficulty, discrimination, and guessing (low level test takers doing well on a difficult item by chance) parameters (Bachman, 2004). Applications of 2-parameter and 3-parameter IRT models are largely restricted to the study of dichotomous items where there are two possible scores (McNamara, 1996). In the current research, 1-parameter IRT model (unidimensional Rasch model) is explored since all C-test texts are assumed to

measure the same underlying trait and polytomous C-test scores are used. This model can be viewed as extended logistic regression which includes a random effect for examinee ability and fixed effects for item difficulty. The model parameters are therefore estimated on the logistic (log-odds) scale.

There are also different Rasch models for different types of data. These models involve the Basic Rasch model for dichotomous data (i.e., multiple-choice questions scored as correct or incorrect), Rating Scale Model (RSM) or Partial Credit Model (PCM) for polytomous data (i.e., Likert-type questions), and Multi-Faceted Rasch Model for data involving ratings mediated by raters (i.e., oral interviews) (McNamara, 1996). Since C-test texts are polytomous (i.e., taking a value between 0 and 20), RSM or PCM is considered the most appropriate. RSM (Andrich, 1978) converts raw scores into true interval scores known as logits. Hence, it assumes that within each C-test, each text item is equally difficult and has an equal probability of being completed within 0-20 points scale. On the other hand, PCM (Masters, 1982) does the raw score conversion individually for each text without assuming a common rating scale for all texts. Therefore, it does not presuppose that each C-test text should be completed within the same 0-20 scale. In other words, it assumes that texts have different levels of difficulty and the scoring scale for each text is different. Compared to PCM, RSM is considered more useful for smaller data sets under 100 and more widely used in the context of C-test texts (i.e., Eckes, 2006, 2007, 2011; Norris, 2006, 2018). The reason why PCM requires a greater sample size is that when each item has a different rating scale, the required sample size increases considering at least 10 observations per category (Linacre, 2012; Linacre, 2017a). Nevertheless, both RSM and PCM were shown to perform quite similar in terms of the model data fit, reliability, and discrimination although there were differences in item difficulty across

models (Baghaei, 2010). In this study, RSM was preferred considering that the sample size is under 100 and assuming that all texts taps into the same construct of language proficiency. According to the RSM, the probability ($p_{ik}(\theta_{n)} = p(x_{in} = k|\theta_n)$) that the test taker *n* with ability θn will achieve a score of *k (k = 0….,m)* on item *i* is formulated as the following (Eckes, 2011):

$$exp \sum_{j=0}^{k} [\theta n - (\beta_i + \tau_j)]$$

$$p_{ik}(\theta_{n)} = p(x_{in} = k|\theta_n) = \frac{}{\sum_{r=0}^{m} \boldsymbol{exp} \sum_{j=0}^{r} [\theta \boldsymbol{n} - (\beta_{\boldsymbol{i}} + \tau_{\boldsymbol{j}})]}$$

4.5.2.2 Interpreting Rasch Output

This section explains and defines key item and test taker statistics produced by Rasch analysis as well as the criteria used to evaluate these statistics.

4.5.2.2.1 Test Taker Statistics

The following test taker statistics are explained in this section: examinee fit indices and examinee separation indices. First, examinee fit indices (infit and outfit mean-squares) help to find inconsistencies within test takers by comparing their observed and expected scores taking into consideration the scores of other test takers. For example, if a test taker left some easy items empty while doing exceptionally well on difficult items, that test taker might be identified as an outlier. Outliers are shown by infit and outfit mean-squares (MnSq) bigger than 2.0. Linacre (2018) said MnSq values indicate the size of the randomness in the data and if they are bigger than 2.0, they distort the model. Outlier examinees should be inspected by examining their

actual responses to all items, then the problematic items might need revision or outlier examines might be removed.

Second, examinee separation indices (separation, strata, and reliability) show to what extent the test was able to reliably separate test takers into statistically distinct ability levels. Separation states the number of statistically distinct ability levels when very low and very high scores are considered as measurement error. Strata also states the number of statistically distinct ability levels, but it is preferred over separation when very low and very high scores are considered as "extreme true levels of performance" (Linacre, 2018, p. 237). In other words, separation is preferred if the distribution is normal, and strata is preferred when the distribution is heavy tailed. (Linacre, 2018). Therefore, it can be said that strata is more relevant when extreme scores are included in the analysis as extreme true levels. Examinee (separation) reliability is the "Rasch equivalent of the KR-20 or Cronbach Alpha" (Linacre, 2018, p. 327), and it is reported as test reliability. It relates to the probability that the test can distinguish among high scorers and low scorers.

Regarding the interpretation of examinee separation and reliability, if separation is lower than 2 and reliability is lower than 0.8, it means that the test may not be sensitive enough to separate high levels from low levels and more test items are required (Linacre, 2017b). Linacre (2017b, p. 638) gives the following tentative guidelines for examinee reliability to decide whether the test separates the sample of test takers into enough levels for its purpose: "0.9 = 3 or 4 levels, 0.8 = 2 or 3 levels, 0.5 = 1 or 2 levels".

4.5.2.2.2 Item Statistics

The following item statistics are explained in this section: item fit indices, item correlations with the overall test, item discrimination values, standard error values,

and item measure values. First, item fit indices (infit and outfit MnSq) show how much each item fits with the overall pattern expected by the measurement model. There are three cases regarding item fit: 1) items showing good fit which allow for normal variation between observed and expected scores, 2) underfit items which show excess variation (noise) than expected, 3) overfit items which show too little variation and depict a deterministic rather than a probabilistic pattern (McNamara, 1996). The productive measurement range for Infit and outfit MnSq values is between 0.5 and 1.5 while a more stringent range is between 0.7 and 1.3 (Linacre, 2018). Underfit items are shown by MnSq values higher than 1.5, and overfit items by MnSq values lower than 0.5. These misfit items might indicate problems with test content or construct such as they might be poorly written, or they might be very different from other items in the test. Therefore, they should be revised and discarded from the test if necessary.

Outfit MnSq indicates that person and item measures are unexpectedly very different such as a low-level test taker answering a difficult item correctly. It is more an indication of guessing or mistakes on easy items caused by outliers. On the other hand, infit MnSq shows that person and item measures are unexpectedly highly similar such as all low-level learners doing an easy item correctly without any occasional violations. If infit MnSq is not within the productive measurement range, it is more problematic than outfit MnSq because it is caused by common response patterns of test takers (i.e., 'all' low level learners doing an easy item correctly while we expect to see some noise) and harder to diagnose the reason (Linacre, 2018).

The point-biserial correlation coefficient (Rpbi) is the Rasch version of the Pearson correlation (Linacre, 2018). It shows the strength of the relation between an item and the overall test. It is useful to ensure that all items are consistent with each other. It ranges between 0.0 and 1.0, and a minimum value of 0.8 is expected (Norris,

2018). Item discrimination values reported by 1-parameter Rasch model are not parameter estimates since it involves only one parameter of item difficulty contrary to 2-parameter Rasch model which involves both item difficulty and item discrimination parameters (see section 4.5.2.1). However, 1-parameter Rasch model yields results of item discrimination as descriptive statistics. It assumes that all item discrimination values are equal to 1.0 to fit the model; however, since item discriminations are never exactly equal, it can provide "an estimate of those discriminations post-hoc (as a type of fit statistic)" (Linacre, 2017b, p. 135). Items with discrimination values closer to 1 are considered to fit the model the most, and values ranging from 0.5 to 1.5 indicate reasonable fit to the model (Linacre, 2017b).

Item measure values show each item's difficulty estimates in logits. Thus, they allow us to see how much more difficult each item is from one another on the logistic scale. The average difficulty of items is set to zero, and positive values above zero indicate more difficult items while negative values below zero indicate easier items (McNamara, 1996). The error associated with item measures are also reported as standard errors alongside. The lower the standard error is, the more precisely estimated the item measures are. The same logistic scale is used for both examinee ability and item difficulty. Thus, Rasch analysis allows us to examine the relationship between item difficulty and examinee ability on the same logistic scale, and we can see whether an item is easy or difficult for the sample of test takers. Measures of item difficulty and examinee ability are graphically presented on the same logistic scale in item-person (or Wright) maps.

4.5.2.3 Differential Item Functioning

Differential Item Functioning (DIF) can also be used to investigate whether test items function differently for different groups of test takers due to a construct-irrelevant

factor (i.e., factors that are irrelevant to language proficiency but may have an impact on test scores) such as gender. If an item shows different probabilities of success for two persons of the same ability level (determined by person ability measure on the logit scale) but with different values on the test-irrelevant factor, that item is said to display bias with respect to that factor. There are two criteria to identify items that have bias: (1) DIF contrast, which means the difference of difficulty of an item between groups, is bigger than or equal to 0.50, (2) the probability value, which means the chance of observing DIF contrast by chance, is smaller than or equal to .05. (Winsteps Tutorial15).

Even if there is an item bias, DIF analysis does not explain why an item benefits a group over another. To understand the reason of item bias, content analysis should be conducted on the biased item. If necessary, that item should be replaced or omitted.

4.5.3 Thematic Analysis

The present research utilises thematic analysis which is a process for "identifying, analysing, and reporting patterns (themes) within data" (Braun & Clarke, 2006, p. 79). Thematic analysis is not tied to a particular theoretical framework. The reason to choose this analysis approach is to have flexibility to reflect and make active choices while coding the data. Braun and Clarke (2006) provided a step-by-step approach as a guideline to conduct thematic analysis. However, they emphasized that these steps are not rules and can be applied flexibly to fit the specific study aims.

The data will be analysed by following the steps of thematic analysis recommended by Braun and Clarke (2006) as seen in Figure 8. The analyses will be

15 https://www.winsteps.com/a/winsteps-tutorial-4.pdf

conducted separately for different groups of participants (i.e., learners and instructors).

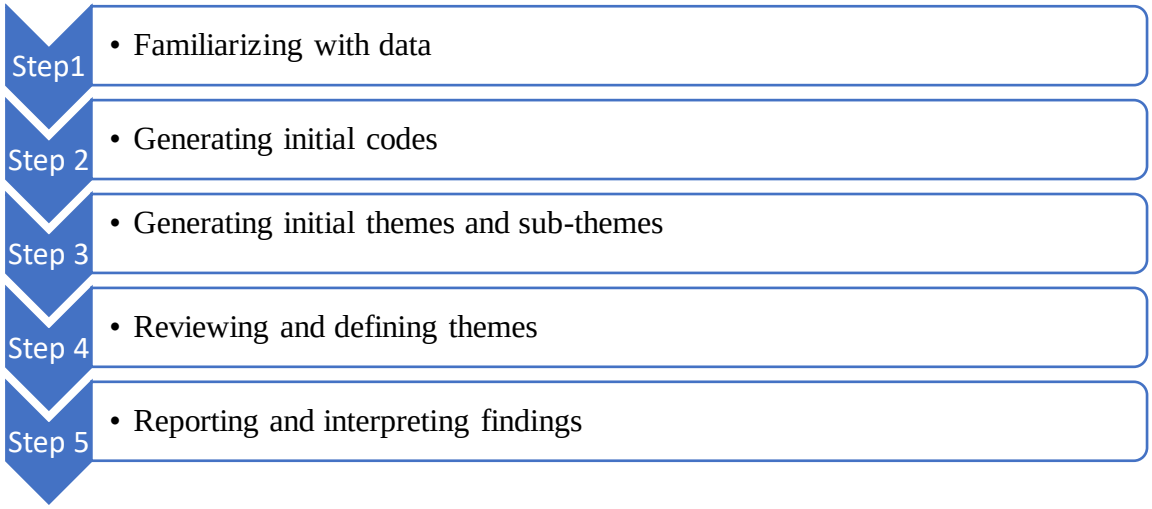


Figure 8. Steps of Thematic Analysis (Braun & Clarke, 2006)

4.5.3.1 Familiarizing with data

First, interviews will be transcribed manually. Then, the interviews, which are conducted in Turkish upon participants' requests, will be translated into English in order to make categorization process easier later. Manual transcription and translation will pave the way for familiarizing with data. All interviews and open-ended survey answers will be organized, imported into the NVivo 12 software and read several times. During this stage, the points which seem interesting and important will be highlighted.

4.5.3.2 Generating initial codes

After familiarizing with data, coding will be done with the help of NVivo by organizing codes and quotations. Initial codes (nodes on NVivo) will be generated deductively by searching interesting aspects of the data related to the relevant evaluation questions. Quotations from the data will be categorized under relevant codes.

4.5.3.3 Generating initial themes and sub-themes

After generating codes, themes and sub-themes will be generated separately for each group of participants (i.e., learners, researchers, instructors) by gathering and interpreting codes under broader terms. For this process, the common features and repeated patterns across codes will be examined, and codes will be sorted into broader themes (see results sections 6.7.4 and 7.7.4 of validation studies for thematic maps).

4.5.3.4 Reviewing and defining themes

After the initial themes are generated, they will be reviewed to examine their relationship with each other and the relevant evaluation questions. Following this, the boundaries and specifics of each theme will be defined considering what distinguishes each theme. The data will be read again considering the generated themes to identify whether there are any codes and additional themes that are missed during the initial coding.

4.5.3.5 Reporting and interpreting themes

The report will be centred around the main themes generated from each group of participants. Vivid quotes from the data which are stored under nodes on NVivo will be included and interpreted under relevant themes.

4.6 Ethical Considerations

Ethical approval was received before conducting this research (see Appendix 1 for the confirmation of ethical approval and the ethics form). The researcher of this study will assign numbers to test participants to ensure that no data can be traced to participants' identities. The data of the participants will be stored in a password-protected personal computer and the identify of participants will not be shared with third parties. This research will involve adult Turkish L2 learners as well as Turkish L2 instructors and researchers. It will exclude children and adolescents under 16. Participants will be

asked to provide their informed consent before participating the research. The consent form will involve information about research purpose, benefits and nonexistence of any associated risks, test instructions, data protection and anonymity, participants' right to withdraw from the research any time, and the contact information of the researcher as well as her supervisors. All participants will be provided with a £5 (or $5) Starbucks gift card or £25 Idefix gift card upon completion of the study to compensate for their time. There will also be a £50 prize draw for the interviewees who volunteer to participate for the semi-structured interviews Any ethical issues are not perceived with this reward since it is a reasonable amount and supposed to encourage higher level of response. Thus, a more representative sample of the population is expected to be achieved. Regarding interviews, member checking will be conducted, that is, transcriptions will be sent to interviewees, and they will be asked to revise the transcriptions to ensure accuracy and transparency of the interviews.

CHAPTER 5: TEST DEVELOPMENT

5.1 Introduction

This chapter reports on the development and initial investigation of a new Turkish C-test (see section 2.3.2 for a general description of C-tests). The existing few studies on Turkish C-test focused on Turkish-German bilingual pupils and did not explore the uses of Turkish C-test for adult Turkish L2 learners (see section 3.3.2 for details of existing Turkish C-test research). This chapter details the test development process stage by stage with language specific factors and investigates whether C-test deletion is applicable to Turkish language when the test is administered to adult Turkish L2 learners. First, it explains how texts were selected and words were deleted considering the typological structure of Turkish explained in section 3.3. Then, it moves onto pilot testing with Turkish native speakers. Following this, it explains the first trial administration with Turkish L2 learners to evaluate the feasibility of the Turkish C-test with this population. Based on the findings of this initial investigation described in this chapter, the test is revised, and the best functioning texts are chosen for the final test version before validating its uses for SLA research in Chapter 6 and educational screening purposes in Chapter 7.

5.2 Text Selection

Following the general principles of C-test development suggested by Klein-Braley (1997) as outlined in section 2.3.2.1, authentic texts were selected from the mainstream media aimed at Turkish native speakers and two texts were adapted from a commercial Turkish textbook (Öztopçu, 2009) to capture lower levels of proficiency. Sources of mainstream media involved newspapers (i.e., editorial column part, nationwide news), journals, websites of organizations (i.e., school, health organization), Vikipedi (Wikipedia for Turkish), blogs, and a graded Turkish reader

book consisting of authentic graded texts according to the Interagency Language Roundtable (ILR) scale, which consists of a set of proficiency level descriptions. The chosen texts of the C-test varied in their estimated level of difficulty. They were assigned a level on the ILR (1, 1+, 2, 2+, 3) by the researcher and a Turkish ILR level rating expert considering text mode, text type, vocabulary, content, and structural forms. Level assignments by the researcher and ILR rating expert were the same most of the time with one plus (+) level of difference in three texts. Any disagreement regarding the level of texts was resolved through discussion. Subsequently, two Turkish instructors rated the texts according to the curricular levels of Turkish language instruction found at universities in the USA. There was a moderate agreement between language instructors since they had one level of difference in their level assignments on five texts. Note that Turkish instructors were from different institutions and they might have contingently been referring to different curriculums during this process.

Figure 9 below, which was taken from Dirgin (2014), summarises the text modes on the main ILR levels. For example, ILR level 2 texts are instructive and informative stating facts, news, or reports. Note that, in addition to the main levels seen in the Figure 9, there are also plus levels (i.e., 1+, 2+) which meet the basic level requirements of the next main level but fail to meet all the criteria to reach the mastery of that level.

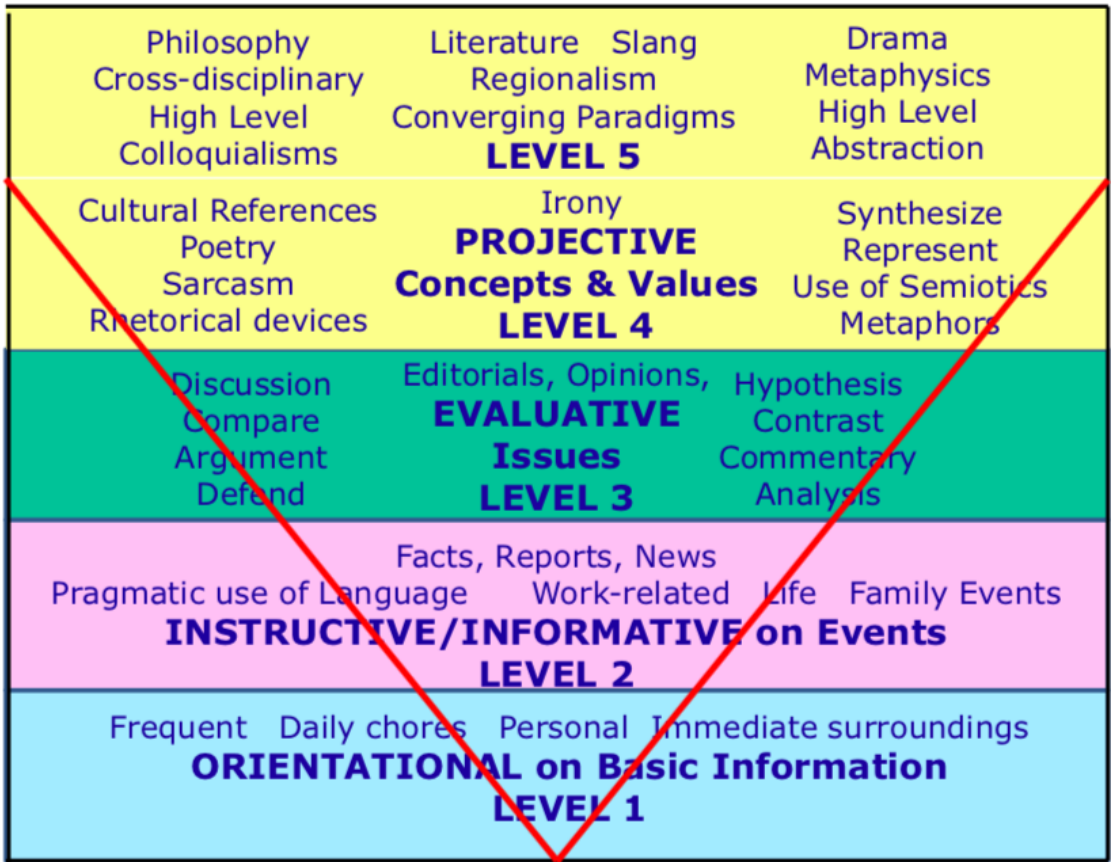


Figure 9. Text modes on ILR scale

Texts over ILR level 3 were found to be inappropriate for C-test construction since they were too abstract, literary, or technical to be reconstructed considering Klein-Braley (1997)'s guidelines for developing C-tests (see section 2.3.2.1). They also required a longer context than short C-test texts to fill in the gaps. Nevertheless, the analyses of this chapter showed that texts over ILR level 3 would be necessary if the aim was to distinguish among higher level learners as well, and thus a new text was included in the test after the analyses (see section 5.6 for details).

Since authentic texts at ILR level 1 lack discourse (use of language in a context) features, such as cohesive devices (i.e., however, because), level 1 texts were chosen from language teaching materials aimed at Turkish L2 learners to ensure that they would be appropriate for lower level learners aligning with Lee-Ellis (2009). Overall, the difficulty of the Turkish C-test was mainly determined by the paragraph

difficulty using ILR reading skill level description before applying the word deletion because text-level scores (C-test texts as superitems) was used in analysing data rather than individual gap-level (gaps as items) scores. This is in line with Khoshdel et al (2016)'s suggestion to focus on text-level characteristics rather than gap-level characteristics in determining text difficulty since text-level characteristics were found to explain most of the variance in text difficulty (see section 2.3.2.2 for details).

At the end of the text selection, a total of 18 texts covering different presumed levels of difficulty were collected, following Grotjahn (1987) who suggested that a researcher should begin with at least twice as many C-test texts as the actual test will consist of, as some of the texts may be excluded due to statistical properties and other factors (i.e., low accuracy percentage with native speakers). To the extent possible, the texts were neutral in content, appropriate for the target group (adult Turkish L2 learners with a wide range of proficiency levels) and did not contain any technical vocabulary or knowledge.

5.3 Word deletion strategy

The C-test texts were prepared according to general C-test deletion principles (Klein-Braley, 1997), which were detailed in section 2.3.2.1. This section first summarises the challenges that were faced in applying the second-half deletion rule to Turkish due to its morphological structure (see sections 3.3 and 3.3.1 for details about Turkish morphology and the resulting challenges). Then, it explains how these challenges were addressed.

First, when the second-half deletion method was applied to the chosen Turkish texts, most of the grammatical information was deleted because of the agglutination (forming complex words by stringing morphemes together) in Turkish, which is illustrated in the example below.

Example: Motivasyon + unuz
Motivation second person marker

Motivas_____ (Turkish)

Your motiv_____ (English)

When the word *motivasyonunuz* (your motivation) is deleted, both the second person marker -*unuz* (your), which is attached to the word root, and some letters of the word root *motivasyon* (motivation) are deleted. On the other hand, in the English equivalent of this word, only half of the word "motivation" is deleted while the grammatical word "your" is left untouched, since it is an independent function (grammatical) word.

In order to overcome this challenge, it was carefully considered whether there were enough contextual clues in the texts to help learners fill in the blanks. Turkish is a pro-drop language, however, where some pronouns may be omitted and inferred from the context. In the above example, the subject pronoun *sizin* (your) is omitted and inferred from the second person marker -*unuz* attached to the word root. However, when -*unuz* is deleted with the second-half deletion method, inserting the subject pronoun *sizin* might be necessary.

Another example below shows the potential ambiguity regarding the pro-drop pronouns used in the sentences.

Ev + e *gel + di + ği + ni* *gör + dü + m.*
noun – dative verb – past tense m. – verbal adjective – personal m. verb – past simple – personal m.

I saw that you came home.

In the sentence above, neither the subject *ben* (I) nor the subject of the nominal clause *senin* (you) is explicitly stated as an individual word. Rather, they are marked as suffixes –*ni* and -*m* at the end of the verbs. Thus, when these suffixes are deleted with the second-half deletion method, it would likely be confusing how to complete them. Therefore, some pronouns were added to the texts, in order to remove

any possible confusion, as seen in the example below where the subject of the nominal *senin* (you) is included:

Senin	*eve*	*geldiğini*	*gördüm.*
You	*to home*	*came*	*I saw*

One additional distinction for the Turkish C-test was that each passage had only 20 deleted words, rather than 25, which is more typically observed in C-tests of European languages. Since Turkish texts contain relatively little redundant information, 20 gaps would provide greater context with additional complete sentences. Thus, sentences that came after the 20th deletion were left intact, and the text continued to a natural end. Note that there was no uniformity and justification regarding the number of deleted words in previous Turkish C-test studies. Some preferred 20 deletions per texts (Daller et al, 2002; Sağın-Şimşek, 2006) while others preferred 25 deletions (Baur & Meder, 1994; Caprez, Gönç, 2006).

Another accommodation for deletions in Turkish texts that the word *ve* (and) was left untouched since this item had an item discrimination value at .00 and item facility value at 1.00 in a previous pilot study (i.e., it is almost impossible not to accurately fill in the second half of this word when it is deleted) (see section 4.5.1 for explanations of item facility and discrimination). In addition, the general additive operator *–de/da* (also) was omitted in deletions, as it did not add to the meaning of the sentences. Regarding compound words (words formed by combining two or more words), the second half of the second word in the compound was deleted. For two words, one more letter was left standing to avoid possible alternative answers, as the given context was not specific enough to reconstruct those words as illustrated in the example below.

Öğre(t)_____ sevmek de motiv_____ artıran b_____ faktördür[16].

The first word of the sentence could be *öğretmeni* (teacher) or *öğrenmeyi* (learning), both of which have the same word root and work well in the given context. Therefore, one more letter -t was left standing (shown in parenthesis).

After the deletion of words in the texts, three texts were eliminated, as the content was deemed insufficient to reconstruct the words appropriately. The remaining fifteen passages were considered appropriate. Finally, unlike in operational C-tests, the texts were ordered in a mixed way rather than following an increasing level of difficulty, to ensure that test takers did their best to try all the texts equally and did not give up when they came to the hardest last texts considering the large number of texts compared to usual C-tests with 4 to 6 texts (see section 5.7 for a discussion of these decisions made during test development).

5.4 Pilot testing with native speakers

Pilot testing was carried out with a group of native speakers of Turkish (N=19) as suggested by Klein-Braley (1985). In comparison to prior Turkish C-test research, the present research tried to establish the appropriateness of difficulty of C-test texts by recruiting native speakers as control group and eliminating the texts where a certain level of accuracy was not achieved by native speakers before administrating the test to language learners. As detailed in section 2.2.1, according to Hulstijn's language proficiency model, native speakers vary in HLC, where they might fall behind non-native speakers. Therefore, only native speakers with a certain level of literacy and educational background (i.e., adults who have completed at least high school education) were recruited for the pilot study.

[16] Liking a teacher is also a factor that increases motivation

The ages of participating native speakers ranged between 19 and 30 years. First, five native speakers (graduate students at a US university) took the fifteen-text C-test as well as a very simple background questionnaire and C-test questionnaire in Turkish. Following this first pilot testing, the number of texts was reduced to thirteen. One of the texts was eliminated as the context provided was not enough for the native speakers to respond correctly. The other involved more spoken language (presented in a written format), inverted sentences due to free word order, and a shift of pronouns that was found to be difficult. In the next round, ten more native speakers, who were more diverse in terms of educational background took the thirteen-text C-test. Although these participants had a certain level of literacy, they differed in their completed level of education (i.e., high school, university) and their area of study such as engineering and politics. Based on their answers, one more letter was left standing in some of the deleted words to reduce the number of alternative answers. Two more texts, where native speakers were not able to reconstruct at least 90% of the missing items correctly, were eliminated. Native speakers commented that these texts involved uncommon collocations as well as very long sentences and clauses. For example, the sentence given below, which is taken from one of the eliminated texts, is long and contains a very long subject partly because of the modifying attributive verb (verb which modifies a noun) *gelen* (coming).

Büyük b___ bölümü 1960'larda çal___ amacıyla Avr___ ülkelerine ge___ Türkiye ***köken***___ yaşadıkları ülke____ kısmen ya___ asırı ger___ bırakırken bulun___ toplumların ayrı___ parçası hal___ geldiler.

While *the Turkey* ***originated people****, many of whom came to European countries in order to work in the 1960s*, left behind nearly half a century in the

countries where they live, they became an inseparable part of the societies to which they belong.

The italicized part states the subject of the sentence and the bolded word shows the subject root. While the subject root is (originated people) given at the beginning of the sentence in the English version, it is at the end of the italicized subjective relative clause in the form of a declined adjective *kökenliler*, which takes the plural marker -*ler* (s) to indicate people and serves as a noun in the Turkish sentence. This was found to make the other preceding words harder to complete.

Finally, four more native speakers took the eleven-text C-test, and they were able to complete all the texts with at least 90% accuracy rate[17]. Table 6 shows the average accuracy percentage for native speakers on each text.

Table 6. Average accuracy of native speaker completion for 11 C-test texts

Text 1	**Text 2**	**Text 3**	**Text 4**	**Text 5**	**Text 6**	**Text 7**	**Text 8**	**Text 9**	**Text 10**	**Text 11**
100	100	98.75	97.75	97.15	95.75	96.25	95.40	91.75	97.5	92.45

Details regarding the ILR level and content of these remaining final 11 texts are provided in Table 7.

Table 7. Levels and content of the 11 C-test texts

Text	**ILR Level**	**Topic**	**Characteristics**	**Source**	**Excluded after analysis**
1	1	Locations	Very basic sentences with "there is/there are" structure Familiar words, cognates	Created based on commercial textbooks	No
2	1	Daily life routine	Short, simple sentences with present continuous Concrete and high frequency words	Created based on commercial textbooks	Yes

[17] Following the analyses, the threshold level was adjusted for 80% accuracy level (see section 5.7 for discussion)

3	1	A Danish person in Turkey	Simple sentences with present continuous Concrete words, some cohesive devices	Graded Turkish reader book (authentic texts)	No
4	1+	Description of a Turkish City	Simple sentences with relative clauses Informative social purpose	Adapted from an airline website	No
5	1+	Biography of a Turkish Singer	Simple sentences with past simple Concrete but less frequent words	Adapted from Wikipedia	Yes
6	1+	The advertisement of a School	Simple and compound sentences. Concrete words, a few abstract words	A School website	Yes
7	2	Student success	Cause and effect relations Factual information	A health organization website	No
8	2	Turkish Language Education	Compound sentences with passive Abstract and concrete lexicon	Journal	Yes
9	2+	Relation between taste and smell	Conditionals and negations Topic specific vocabulary	Newspaper	No
10	2+	Production and motivation	Compound and complex sentences (subordination, embedding) Low frequency abstract words	Newspaper	Yes
11	3	Turkish science Women	Social issue, abstract topic General report Evaluative statements	Newspaper editorial column	Yes

Based on the pilot study with 19 native speakers, alternative and acceptable answers for deleted items were identified and included in the answer key. Instructions were also improved, and changes were made regarding the coherence and cohesion of the texts.

5.5 Trial administration with Turkish L2 learners

5.5.1 Participants

Thirty-seven Turkish L2 learners participated in the trial administration. They consisted of undergraduate and graduate students studying Turkish as a foreign language at three different US universities. Of these, ten participants were enrolled in an intensive beginner Turkish course, seven in an intermediate Turkish course, and eleven in an advanced Turkish course. All beginner level students had been learning Turkish only for 3 months when they took the test. Nine of the participants had completed a full three-year Turkish instruction program at the university level. Turkish language instruction at US universities typically lasts three years (beginner I-II, intermediate I-II, advanced I-II) while a few universities sometimes provide students with elective Turkish courses (i.e., contemporary Turkish composition, media and translation) after the completion of three-year instruction.

The ages of the participants ranged between 18 and 77 years, and there was one untypical student at the age of 77 among beginner-level students. There were 23 females and 14 males. Participants varied in student status from freshman to the Ph.D. level: there were 11 undergraduate, 18 M.A., and, 8 Ph.D. students in total. The participants also had various first languages. There was 1 Hebrew-English bilingual, and 1 L1 speaker of each of the following languages: Arabic, French, Hebrew, Mandarin, Polish, Russian, Persian, and Serbian. However, the primary L1 was English with 26 English L1 speakers. Finally, two of the participants self-identified as Turkish heritage speakers. Nevertheless, they were not among the students who had top scores, and their average scores were 10.36 and 7.82 out of 20.

5.5.2 Instruments

5.5.2.1 Background Questionnaire

A background questionnaire in English (see Appendix 2) was administered to participants prior to the administration of the C-test to gather information about test taker characteristics. The background questionnaire asked participants to report on a range of demographics (age, gender, L1, year in college, major), their language background and use (institutional level of Turkish proficiency, list of Turkish classes taken, length of studying Turkish, age of first exposure to Turkish, time spent in Turkish speaking country, weekly use of Turkish outside class, any other L2s, any family member speaking Turkish, any taken Turkish proficiency test result). Participants were also asked to state their self-perceived overall proficiency as well as proficiency in four main skills (reading, listening, writing, speaking) in Turkish (giving a score out of 5 on Likert Scale as 1 being "beginner" and 5 being "very advanced").

5.5.2.2 Turkish C-test

The Turkish C-test involved instructions at the beginning of 11 texts (see Appendix 3). Instructions were detailed and comprised information about the general format of the test with a practice item. Participants were advised what to focus on and what kind of strategies they can follow to solve C-Test items. They were warned about not to use any dictionary or external aids and be careful about spelling since 100% accuracy is required.

5.5.2.3 Post-test Questionnaire

Following completion of the C-test, a post-test questionnaire in English was administered to participants to collect evidence about the face validity of the Turkish C-test (see Appendix 4). Participants were asked yes/no questions and open-ended

questions about the clarity of the test and instructions, difficulty of the test, and familiarity with any of the texts. They were also asked to rate the difficulty level of texts on a 5-point Likert-scale.

5.5.3 Data Collection

Six Turkish language instructors at several US universities were contacted to request administering the Turkish C-test to students in their classrooms. Some Turkish language learners were also contacted directly through e-mail. Upon the approval of three instructors at three different universities, the C-test was administered in paper and pencil format during students' usual class hours. A few students self-administered the test at home without any consultation to outside resources.

Participants were provided with detailed written instructions. The test duration was not timed to ensure that test takers try as much as they can to do the test. However, it was recommended that they do not spend more than one hour to do it. Furthermore, they were asked to write their test start and end time in order to find the ideal test duration they require to solve the questions. The mean test duration was found to be forty-two minutes, and only three students out of thirty-seven spent more than one hour on the full test. All participants were awarded $10 Starbucks gift card upon completion of the test.

5.5.4 C-test scoring

Dichotomous scoring was applied to rate the C-test. Each answer was given 0 or 1 depending on complete accuracy; alternative answers were accepted only if the sentence was semantically acceptable and there was no change in meaning. As one the aims of the study was to investigate whether the C-test deletion method is applicable to the Turkish language, a zero score was given where there were any spelling

mistakes or morphological errors, such as missing case markers, in order to yield more objective results.

5.5.5 Analyses

C-test data were analysed using a Rasch Model approach in order to investigate item difficulty, item discrimination, test reliability, and the relation between examinee ability and item difficulty (see sections 4.5.2.1 and 4.5.2.2 for explanations of Rasch model and these terms). Because each item was dependent on the corresponding text, the Rating Scale Model (RSM) approach was adopted, with each text treated as a 20-point super-item (Norris, 2006). Analyses were conducted using FACETS (Linacre, 1989).

In addition to Rasch analysis, correlational analyses were conducted between C-test scores and program levels (i.e., level of the course being taken at the time of this study) as well as self-perceived proficiency to check the criterion-related validity of the test. Spearman's rho was selected due to ordinal nature of program level and Likert-scale self-ratings. Following this, one-way ANOVA was conducted between the program levels to investigate whether there are differences in C-test scores of learners belonging to different program levels.

5.5.6 Results

This section reports the results to identify the best functioning texts among 11 texts to form a final C-test. First, it details the findings derived from Rasch analysis and then moves onto the correlational results between C-test scores and program levels as well as self-perceived proficiency.

5.5.6.1 Results of Rasch Analysis

Data were analysed using a 2-Facet (Examinees + Items) RSM. Table 8 below shows the key item quality statistics of the 11-text C-test. The explanations of these item

quality statistics are briefly provided below while reporting the results; however see section 4.5.2.2.2 for more detailed explanations.

Table 8. Key Item Quality Statistics for 11-Text C-test

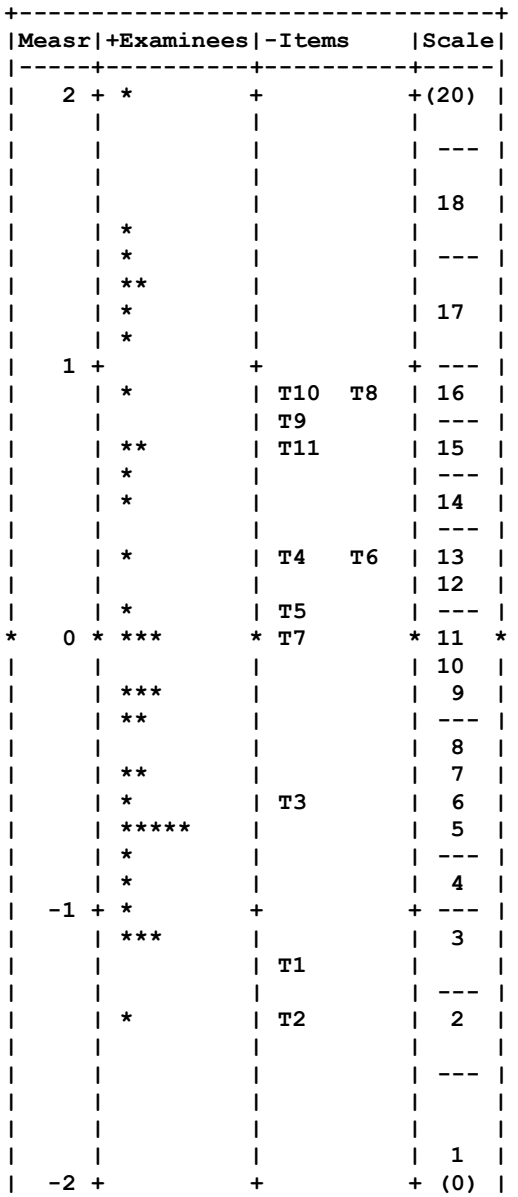
Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.74	1.11	1.05	.91	.09	-1.25
T2	.70	1.00	1.22	1.05	.10	-1.37
T3	.85	1.15	.84	.75	.07	-.59
T4	.96	1.65	.55	.54	.07	.30
T5	.91	1.06	.84	.79	.07	.13
T6	.89	1.12	.96	.86	.07	.34
T7	.91	1.29	.64	.62	.07	.03
T8	.80	.36	1.45	1.81	.09	.85
T9	.91	1.34	.93	.85	.08	.81
T10	.89	1.16	.65	.89	.08	.85
T11	.91	.98	.72	.91	.08	.70

The first item statistics to check are the item fit indices (infit and outfit statistics), which show how much each item fits with the pattern expected by the model. The productive range for item fit indices lies between 0.5 and 1.5 (Linacre, 2018). Point-biserial correlation coefficient (Rpbi) shows the strength of the relationship between an item and the overall test, and a minimum value of 0.8 is expected (Norris, 2018). Discrimination values show how much an item can discriminate among examinees, and the values closer to 1 are considered to fit the model most while values between 0.5 and 1.5 are acceptable (Linacre, 2017b). Finally, measure shows item difficulty in logits and while positive values show more difficult items, negative values show easier items. The error (SE) associated with item measures are also reported alongside, and the lower the standard error is, the more precisely estimated the item measures are.

Considering these threshold values, all texts except T8 had infit/outfit values between 0.5 and 1.5 and discrimination values ranging between approximately .5 and 1.5. All texts except Text 1 and Text2 had point-biserial values higher than 0.8. Item difficulty measures ranged considerably, from -1.37 (T2 is the easiest) to .85 (T10 and

T8 are the most difficult), suggesting a good deal of difference across the items, as intended. Difficulty measures of the texts were similar to their estimated difficulty level according to the ILR scale. All negative values corresponded to texts with ILR level 1 (T1, T2, T3). Texts with ILR level 1+ (T4, T5, T6) had difficulty measures between .13 and .30 while texts with ILR 2 and above (T8, T9, T10, T11) had higher difficulty measures ranging between .70 and .85 with the exception of T7 which was found to have some repetitive words.

The relationship between item difficulty and examinee ability was also examined through item-examinee map as seen on Figure 10 below.



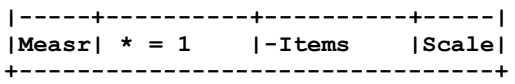


Figure 10. 11-text C-test item-examinee map

As illustrated on Figure 10, examinees and items were distributed relatively well, and to somewhat comparable degrees. Note that examinees are more able, and texts become more difficult, from the bottom to the top. Note also that a handful of examinees exhibited higher ability measures than the most difficult of the 11 texts, which indicated that one or two more difficult texts might be needed. Nevertheless, a majority of examinees fell within the corresponding range of text difficulties. There was also a generally expected relationship between the estimated difficulty of texts and their resulting difficulty measures (i.e., lower numbered texts fall towards the bottom of the figure, while higher numbered texts occur at the higher or more difficult end of the spectrum). While the measure column shows the IRT level difficulty ranging between -2 and +2, the scale shows the score range for the C-test ranging between 0 and 20.

Overall, the 11-text C-test was capable of identifying some 5 levels of examinee proficiency levels with a high reliability (examinee separation = 5.48; strata = 7.64; separation reliability= .97). These examinee separation indices were in line with Linacre's (2017b) guidelines saying that if separation is higher than 2, and reliability is higher than .80, the test is sensitive enough to separate high levels from low levels (see section 4.5.2.2.1 for details). Results show that 90.02% of the total variance in C-test scores was explained by the Rasch model.

After this initial analysis, misfitting texts T2 and T8 were removed. Rasch analysis was conducted again with a new 9-text C-test. The variance accounted for by the new set of texts slightly improved from 90.02% to 90.53%, and separation increased slightly to indicate some 5 levels of differentiation among examinees

(separation = 5.63; strata = 7.84; reliability= .97). The distribution of items and examinees is provided in Figure 11.

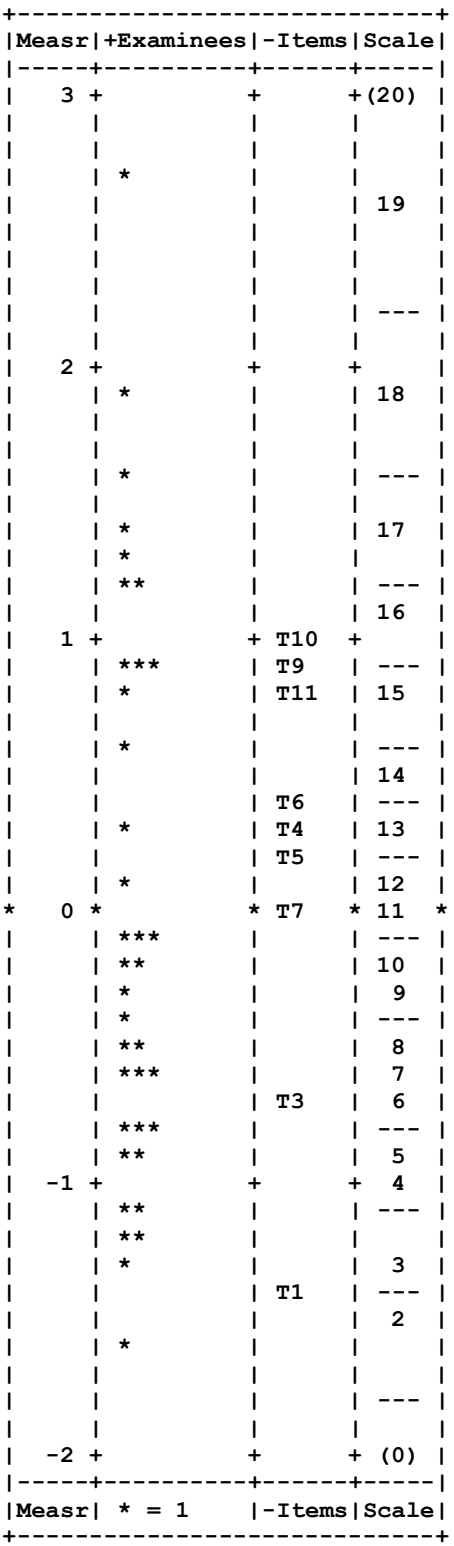


Figure 11. 9-text C-test item-examinee map

In the 9-text C-test, there were no more misfitting items, although T1 and T3 performed marginally as seen on Table 9. However, these texts were retained since they were important for covering lower proficiency examinees.

Table 9. Key Item Quality Statistics for 9-Text C-test

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.73	.87	1.35	1.15	.11	-1.44
T3	.83	1.08	1.04	.87	.08	-.66
T4	.96	1.64	.53	.53	.07	.34
T5	.91	1.11	.85	.78	.07	.15
T6	.90	1.06	.96	.88	.08	.38
T7	.91	1.23	.66	.68	.07	.04
T9	.92	1.28	.90	.83	.08	.92
T10	.89	.98	.80	1.13	.08	.97
T11	.91	.83	.87	.98	.08	.79

Overall, the 9 items were seen to be grouped into approximately 5 different levels of difficulty as seen in Figure 11 and demonstrated by the separation index of 5.63. T1, T3, and T7 each seemed to contribute to a different level among lower ability students. On the other hand, T4, T5, and T6 were within the same grouping of difficulty covering the higher ability students while T10, T9 and T11 were within the same grouping of the most difficult texts addressing the highest-level students. Within these two groupings, the items that had higher discrimination and point-biserial values were chosen to eliminate the redundant texts. For validation study 1 in Chapter 6, to create a Turkish C-test that could be used for research purposes and efficiently administered within a short amount of time, T10, T11, T6, and T5 were eliminated. For validation study 2 in Chapter 7, to create a Turkish C-test that can be used as a screening test for TYS, only T10 and T5 were eliminated with the goal of having two texts per level in order to predict TYS levels.

After eliminating the redundant texts, the Rasch Model analyses were conducted again with the final 5-text C-test that will be used in study 1 and the 7-text C-test that will be used in study 2. The key item quality statistics for the 5-text C-test

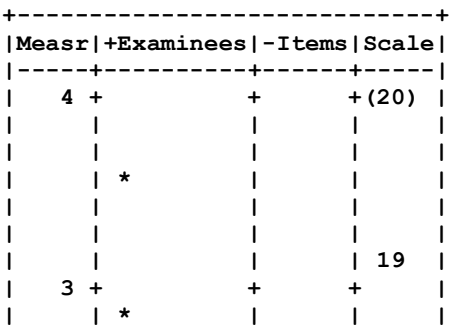
are provided in Table 10. (see Appendix 5 for the results of Rasch Analysis with 7-text C-test).

Table 10. Key Item Quality Statistics for 5-text C-test

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.75	.57	1.31	1.13	.12	-1.76
T3	.84	1.04	1.00	.90	.08	-.82
T4	.93	1.35	.64	.65	.08	.39
T7	.90	1.00	.82	.87	.08	.02
T9	.88	1.20	.79	.81	.10	1.11

All texts except T1 were found to have point-biserial values higher than .80 and discrimination values between 0.5 and 1.5. Infit and outfit statistics of all texts were within the range of 0.5 and 1.5. Of these texts, T1 was the least authentic one which was created based on a dialogue in an elementary level Turkish coursebook for lower level learners. It was important to include lower level texts, though, since beginning students would likely find a Turkish C-test based on more advanced texts overly difficult, both because of the agglutinative structure of Turkish and their general exposure to authentic Turkish input.

The final 5-text C-test item-examinee map is shown in Figure 12. Note that the explained variance increased to 92.47%, while reliability remained very high at .96 despite reducing the test to 5 texts. The 5-text C-test was still able to distinguish reliably across approximately 5 ability levels of examinees (separation = 4.62; strata = 6.50; reliability = .96). The separation indices and reliability were the same for the 7-text C-test (see Appendix 5).



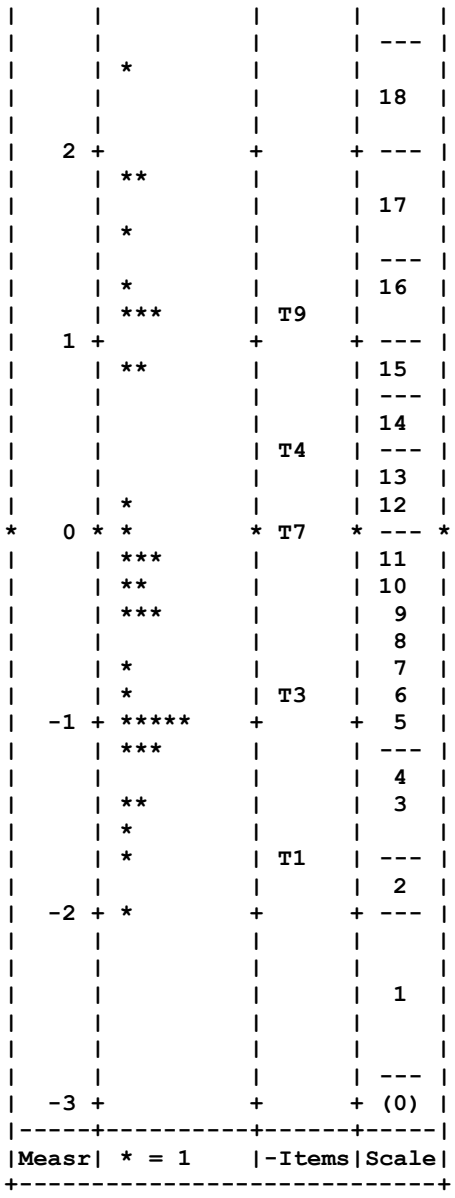


Figure 12. 5-text C-test item-examinee map

As seen in Figure 12, there are three students who might not be covered by the most difficult text (T9). However, when the characteristics of these learners were examined individually, it was found out that all these learners had been learning Turkish for more than four years, and they all had studied Turkish in Turkey after their language instruction in the US was completed. Therefore, it was determined that these learners might be reflecting the highest level of abilities likely to be tested, providing further confirmation of the extent to which the test was accurately measuring the proficiency differences across the spectrum of learner abilities.

Examinee statistics were also examined in addition to item quality statistics (see section 4.5.2.2.1 for details on examinee statistics). One outlier was detected in the patterns of performance (either gave up or didn't try equally hard on all texts). This resulted in examinee infit/outfit statistics higher than 2.0 which is calculated by comparing examinees' observed and expected scores by taking other examinees' scores into consideration. Removing this outlier would slightly improve the results; however, the difference wouldn't be discernible (see Appendix 6 for Rasch analysis results when this outlier is removed). Furthermore, removing the outlier could be interpreted as trying too hard to fit the data to the model given the small sample size (N=37). Therefore, this outlier was not removed.

5.5.6.2 Correlational Analyses

Correlation analyses were conducted to determine the criterion validity of the Turkish C-test with other indicators of language proficiency. Table 11 shows the correlations between 5-text C-test scores and the program level as well as self-assessment of proficiency. The highest correlation was with the program level (ρ=.91), which is higher than the correlations between C-test scores and program levels found in other studies in Norris (2018) (see section 2.3.2.3 for these correlational studies).

Table 11. Correlations between C-test scores and other measures of proficiency (N=37)

	Turkish C-test Measure (5-Text)
Program level	.91
Self-reading	.88
Self-writing	.76
Self-listening	.85
Self-speaking	.85
Self-overall	.86

Note: all correlations statistically significant, $p < .01$

As seen in Table 11, the highest correlated skill was reading, followed by overall, speaking and listening equally, and finally writing. These correlations are also

similar to those found in Norris (2018) (see section 2.3.2.3 for details). Furthermore, they are higher than the average correlation of self-assessment with various indicators of overall proficiency, which is .63, found in Ross (1998).

Following correlational analyses, a one-way ANOVA was used to test for differences in total C test scores between the four program levels (beginner, intermediate, advanced, very advanced). Note that, the number of participants were very close to each other across program levels (level 1=10, level 2=7, level 3=11, level 4=9). A wide range of proficiency was found as seen in Figure 13.

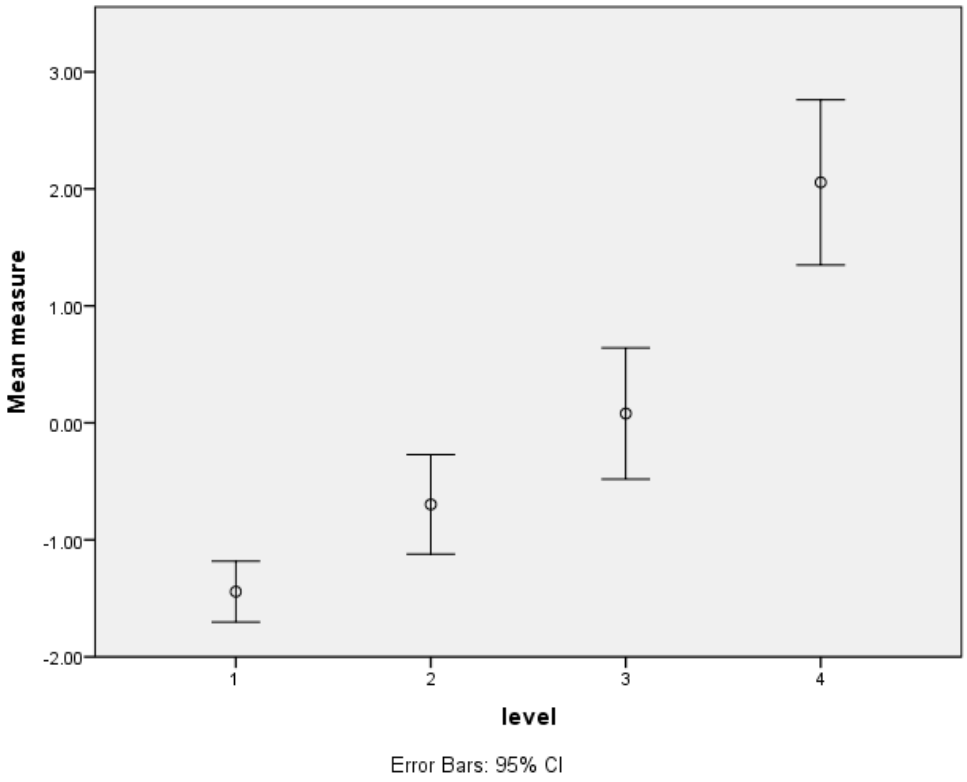


Figure 13. Mean C-test scores by level of Turkish study

There were statistically significant differences between all levels except for 2 and 3. Level 4 students had very high scores compared to others, which explains why there is a group of students beyond the level of the most difficult text. These findings are not overly surprising, however, given that there is typically a good deal of reported heterogeneity within the middle curricular levels, for a variety of reasons.

Indeed, that is one of the main reasons not to trust program level as the only proficiency indicator and to use additional proficiency measures such as a C-test in SLA studies.

5.6 Inclusion of a new text

Following the initial investigation of the Turkish C-test, a new high-level text (ILR 3+) was included among the five best functioning Turkish C-test texts to cover the small group of very advanced level students. The new text was chosen from an academic journal in the social sciences, and its content focused on the relation between cultural venues and folk dances.

The new 6-text C-test was piloted with ten new participants who were native speakers of Turkish. The ages of these participants ranged between 25 and 39 years. Three of them had a bachelor's degree and seven of them had a master's degree. Their area of specialization was diverse such as language and genetics. They were able to complete all the texts except the newly added Text 12 with at least 90% accuracy as shown in Table 12 below. Note that texts are labelled with their original text numbers to be able to compare them across three empirical chapters.

Table 12. Accuracy of native speaker completion for the final 6-text C-test

Text 1	**Text 3**	**Text 4**	**Text 7**	**Text 9**	**Text 12**
96.50	94.00	97.00	91.50	97.50	82.22

In the previous piloting with native speakers, 90% native speaker accuracy was met by eliminating all the texts that fell below this touchstone of 90% accuracy level. However, data analysis revealed that a high-level text was needed for very advanced level learners of Turkish. Therefore, Text 12 was not eliminated in the final version of the test at this stage.

Based on common native speaker mistakes, minor changes were made in Text 12 in order to avoid possible grammatical ambiguity due to a long sentence structure. The example below is a long sentence from Text 12. It contains two verbs and two long subjects. Native speakers could not distinguish the subjects and the verbs of this sentence. Therefore, this long sentence with two different verbs and subjects was divided into two separate sentences and connected through the adverb *bu nedenle* (therefore) as seen in the example below. Another reason for the confusion of native speakers was also due to that the first subject of this sentence (italicized in the example) fell far away from the verb, and its root (bolded in the example) was at the end of a subjective relative clause.

Geleneksel temsillerde öncelikli olan ***mekânlar****,* günümüz koşullarında küresel ve yerel etkilerle değişmiş, kültürel ve mekânsal farklılaşma ve çeşitlilik hızlanmıştır[18].

Geleneksel temsillerde öncelikli olan ***mekânlar****,* günümüz koşullarında küresel ve yerel etkilerle değişmiştir. Bu nedenle kültürel ve mekânsal farklılaşma ve çeşitlilik hızlanmıştır.

Based on native speaker responses, some alternative answers were also added in the answer key due to the morphological productivity of Turkish. Then, the test was ready to be administered to the L2 learners of Turkish.

[18] The venues, which are prioritized in traditional performances, have transformed by the global and local effects of today's conditions, cultural and spatial differentiation and variety have accelerated.

5.7 Discussion

The test development presented in this chapter was the first attempt to design a Turkish C-test which sought to distinguish between adult Turkish L2 learners of different proficiency levels. It was also the first one to specify the development stages of a Turkish C-test step by step with language specific factors. Thus, this chapter provides future researchers and instructors of Turkish with not only a freely accessible Turkish C-test that they can use but also guidance in developing their own Turkish C-test.

While recruiting native speakers for the pilot study, a certain level of literacy and educational background was sought after considering Hulstijn's (2015) language proficiency model according to which native speakers vary in HLC, where they might fall behind non-native speakers. The four texts which native speakers couldn't complete with at least 90% accuracy were eliminated in contrast to the previous Turkish C-test research where ensuring a certain accuracy level with native speakers was not sought before administering the test to learners. Based on native speakers' answers in the C-test and their feedback in the post-test questionnaire, alternative and acceptable answers were identified and included in the answer key, instructions were improved, and changes were made about the coherence and cohesion of the texts. These changes were small and adherence to the original texts was maintained as much as possible to keep the authenticity of texts. However, it is important to emphasize that researchers should be careful while developing a Turkish C-test since some texts might not be suitable for deletion or need adjustments to be constructed as a C-test due to factors such as free word order (depending on which element of the sentence is emphasized), pro-drop structure and extensive agglutination. Therefore, it is essential

to do a pilot study with native speakers to identify any potential ambiguity before administrating the C-test to learners.

Following the test administration to Turkish L2 learners, the number of texts was reduced by eliminating the redundant texts that have unacceptable fit statistics, lower discrimination and point biserial values as well as no unique contribution to the difficulty of the test. The test was reduced to five texts for validation study 1 where it will be used as a research instrument to control the language proficiency of learners in SLA studies (see Chapter 6). In other words, one text per level was kept in each grouping of 5 different levels of difficulty. On the other hand, the test was reduced to seven texts for validation study 2 where it will be used as a screening test for TYS (see Chapter 7). For the test to predict learners' levels on TYS, efforts were made to keep two texts per level. It is also worth noting that there was correspondence between the pre-estimated difficulty level of texts and their resulting difficulty measures with the exception of Text 7, which was found to have several repetitive words. Text 7 was not removed at this stage since it contributed to the overall difficulty level of the test (see Figure 12 above). If future researchers or instructors of Turkish wanted to use the Turkish C-test only with a specific proficiency group of learners (i.e., only beginner levels or only advanced levels), they could choose different combinations of texts among 11 texts.

The elimination of the redundant texts allowed both tests to be more practical and completed in a shorter amount of time while also keeping high reliability values similar to the ones reported in other studies (i.e., Eckes & Grotjahn, 2006; Sigott, 2004) and distinguishing between 5 different proficiency levels. Nevertheless, there were still 3 very high-level learners whose proficiency levels were beyond the highest-level text. Analysing the individual characteristics of these learners closely, it

was found that all these learners had been learning Turkish for more than four years, had studied Turkish in Turkey, and one had a Turkish spouse. Therefore, they are very likely to be representing the highest level of learners. The great difference in the mean measures between Level 4 students and all other levels according to institutional status also supported this inference. Since the aim of the Turkish C-test is not to distinguish among advanced level learners, and it is rather to spread learners across a continuum based on their proficiency, this should not be problematic for this study. Nevertheless, considering the small sample size, a new high-level text (ILR 3 +) was included among the chosen texts to be operationalised in study 1 and study 2.

Texts over the level of ILR 3 were initially considered inappropriate for C-test construction since their topics were too abstract, technical or literary considering Klein-Braley's (1997) suggestion to keep the texts as neutral as possible in content. As the texts get more difficult, they unavoidably involve cultural references and more subject specific knowledge. Nevertheless, learners with proficiency levels over ILR 3 are expected to understand these technical or literary texts which may involve unfamiliar subjects and cultural references (ILR, 1985). Since the results showed that the existing texts were not able cover a small group of very advanced level learners, the new Text 12 at ILR 3+ was included in the test. When Text 12 was piloted with native speakers, their level of accuracy on this text was below the 90% touchstone level, which was previously sought. However, previous literature showed that 90% native speaker accuracy was not always reached in some languages such as Bangla and Turkish (i.e., McKay & Abedin, 2018; Daller et al, 2002). Perhaps, as also suggested by McKay and Abedin (2018), one of the implications of this finding might be that 80% native speaker accuracy level can be taken as the criterion in some languages to allow a wider pool of candidate texts for L2 learners. Another

implication is that for advanced level L2 learners, technical or literary texts involving subject specific knowledge or cultural references might be included in the C-test. The findings of the study 1 and study 2 in the next two chapters will shed more light into this since the newly added Text 12 will be administered to a large sample of Turkish L2 learners.

Developing well-functioning yet lower-level C-test texts was also found challenging, which might be attributed to two reasons. First, lower level texts (i.e. ILR 0+, 1) are typically found in authentic materials such as brochures and menus where the text is loosely organized. However, these kinds of authentic materials are not suitable for C-test structure where contextualization and coherence (i.e. sequence of events, descriptions) are usually relied on in order to fill in the gaps. Second, beginner level learners in this test administration had been learning Turkish only for three months when they took the test. Therefore, it is questionable whether they accurately reflected beginner proficiency levels or might rather be considered as pre-beginner level learners. The findings of study 1 and study 2 will shed more light into the suitability of the Turkish C-test for a more heterogeneous and larger sample of beginner level learners.

Future research should explore the importance of some subjective decisions taken in this study although reasonable justifications were given for these decisions based on the literature review. First of all, text-level characteristics and aggregated scores were used to determine the difficulty of the C-test following Khoshdel et al (2016) and local item independence assumption of IRT analysis. Future research can investigate the effect of both text-level and gap-level characteristics to determine C-test difficulty. Second, 20 words were deleted per text instead of 25 words in contrast to most C-tests in European languages since Turkish texts contain relatively little

redundant information. Researchers can compare Turkish C-tests involving texts with 20 gaps versus 25 gaps and explore its effect on the scores. Third, texts were ordered in a mixed way rather than following an increased level of difficulty to ensure that test takers did their best on all texts. Typically, in operational C-tests which involve a relatively smaller number of texts between 4 and 6, texts are ordered following an increased level of difficulty. Therefore, when the chosen texts are operationalised in chapters 6 and 7, they are ordered according to their level of difficulty from the easiest to the most difficult. Future researchers can compare C-tests with texts ordered in a mixed way versus ordered following an increased level of difficulty. Finally, the Turkish C-test in this initial investigation was a power test without having any time limitations to give everyone an opportunity to do their best and find the ideal test duration. The average time duration was found to be 42 minutes for 11 texts (around 4 minutes per text). In the next two chapters, the Turkish C-test is administered as a speeded test with 5 minutes per text. Future research can also explore whether the time limitations has an impact on learner performance on C-test.

CHAPTER 6: VALIDATION STUDY 1

6.1 Introduction

L2 learners show great variability in their L2 learning background and thus their L2 proficiency (Tremblay, 2011). Since proficiency influences L2 learners' performance in experiments, controlling L2 learners' language proficiency is essential in SLA experimental research, and thus researchers need to control the L2 proficiency of their research participants in a timely and cost-effective way. In the Turkish language, there is a lack of standardized and validated language proficiency tests that are freely accessible to SLA researchers and can be completed within researchers' time constraints (see section 3.2.1 for a detailed discussion of proficiency assessment instruments in Turkish SLA).

The purpose of validation study 1 is to validate the use of the newly developed Turkish C-test (see Chapter 5) to control the general language proficiency of Turkish L2 learners in SLA research studies using Kane's (2006) argument-based approach and make the Turkish C-test publicly available to researchers. Since Kane's (2006) argument-based approach involves an interpretive argument and a validity argument (see section 2.4.4 for details about the argument-based approach), this chapter starts with the interpretive argument which specifies the inferences and assumptions of the test use. Then, the validity argument which involves collecting and evaluating evidence for the assumptions of the interpretive argument follows. This involves describing participants, instruments, data collection procedures, data analysis methods as well as reporting results and discussing whether each assumption is supported or rebutted.

6.2 The Interpretative argument of Validation Study 1

An interpretive argument involves the inferences of the test use/interpretation and the assumptions underlying each inference (see section 2.4.4.1 for details about the interpretive argument). The following interpretive argument on Table 13 states the underlying inferences and assumptions of the suggested test use in validation study 1. It is in line with other studies using an argument-based approach (i.e., Chapelle et. al., 2008; Drackert, 2016; Son, 2018). The table also includes the evaluation questions addressing the suggested assumptions in the same format as Huff et. al. (2008) since involving evaluation questions into the framework is considered more appropriate to show the relation between each assumption and the relevant question.

Table 13. Interpretive argument of validation study 1

Assumptions	Evaluation Questions
Theoretical grounds	
1. The common components of general language proficiency are inclusive of, but not limited to, grammar and lexis.	1. What are the components of general language proficiency?
2. C-tests can quickly assess general language proficiency as demonstrated by a considerable amount of literature.	2. What is the evidence showing that C-test can quickly assess general language proficiency?
Scoring	
3. Text selection and word deletion procedures are appropriate to cover a range of L2 learners in terms of Turkish general proficiency.	3. To what extent does the text selection and word deletion procedures produce a test that can cover a range of Turkish L2 learners?
4. Psychometric characteristics of texts are calculated, and the best functioning 5 texts are chosen for the final test version.	4. Which 5 texts discriminate between Turkish L2 learners of different proficiency levels most accurately and reliably?
5. The C-test distributes test takers along a wide continuum of scores.	5. To what extent does the C-test elicit a wide range of scores?

6. The scoring criteria are appropriate for the test.	6. Are the scoring criteria appropriate?
7. The scoring criteria are applied accurately and consistently.	7. Are the scoring criteria applied accurately and consistently?
Generalization	
8. The C-test texts are internally consistent, and they provide reliable estimates of test takers' L2 abilities.	8. To what extent does the C-test provide reliable estimates of test taker's L2 abilities?
9. The C-test functions consistently for Turkish L2 learners from both UK and USA.	9. Does the C-test produce consistent scores for both US and UK samples?
10. Texts are free of bias against any of the two groups.	10. Are texts free of bias towards UK and USA samples?
11. The sample of observations is large enough to control sampling error.	11. Is the sample of observations large enough to control for sampling error?
Extrapolation	
12. The C-test scores correlate with the variables of Turkish learning history and use derived from the background questionnaire.	12. Are there correlations between C-test scores and Turkish learning history as well as use?
13. The C-test scores correlate with institutional level.	13. Are there correlations between C-test scores and institutional level?
14. The C-test scores correlate with self-perceived proficiency in Turkish.	14. Are there correlations between C-test scores and self-perceived proficiency in Turkish?
Decision	
15. The Turkish C-test scores reflect a certain degree of test takers' general language proficiency. They can be used to control for general proficiency levels of Turkish L2 learners in SLA studies.	15. What are the perceptions of the Turkish C-test stakeholders regarding the usefulness, difficulty, structure, and clarity of the Turkish C-test?
16. The Turkish C-test will enable benchmarking, interpretability, generalization, and replicability across SLA studies in Turkish for the proposed test use.	16. To what extent does the Turkish C-test enable benchmarking, interpretability, generalization, and replicability when it is used to control general language proficiency across SLA studies?

As seen in Table 13, the interpretive argument for validation study 1 includes five inferences: theoretical grounds, scoring, generalization, extrapolation, and decision. The first inference, theoretical grounds, defines the construct of general language proficiency and then provides theoretical justification by connecting the C-test format to the general language proficiency through a considerable amount of existing literature (see Chapter 2).

The second inference, scoring, links Turkish learners' observed performance on the Turkish C-test to their C-test scores which reflect their general language proficiency in Turkish. The assumptions of the scoring inference are based on the test development stage (see Chapter 5 for test development), the sufficiency of the test to elicit a wide range of scores, the selection of the best functioning texts to produce the final 5-text C-test, and the scoring criteria (appropriateness, accuracy, consistency)

The third inference, generalization, links learners' observed C-test scores to their expected scores across C-test texts. The assumptions of the generalization inference are based on the reliability of C-test texts, consistency of C-test scores for learners from both UK and USA, detection of bias for either group of learners, and the sufficiency of the sample size.

The fourth inference, extrapolation, links learners' scores on the C-test to their Turkish level on other indicators of general language proficiency. Its assumptions are based on the correlational studies which explore the relationship between Turkish C-test scores and criterion scores on other measurements.

The fifth inference, decision, links Turkish C-test scores to the intended use of the Turkish C-test. Its assumptions are based on the input of stakeholders, namely SLA researchers of Turkish and Turkish L2 learners, who would use the Turkish C-test.

6.3 Participants

The validation study 1 involved two types of participants who would use the Turkish C-test: Turkish L2 learners and SLA researchers of Turkish. This section presents the demographic information of these participants.

6.3.1 Turkish L2 learners

A total of 85 adult Turkish L2 learners who have learned Turkish in academic language classrooms participated in validation study 1. Among, there were 44 female, 38 male, and 1 other (non-binary) gender. 2 participants preferred not to state their gender. Of these learners, 53 were from North America[19] (N=48 from the USA, N=5 from Canada), and 32 were from the UK. Table 14 shows the distribution of the participants according to the institutional level (the level of language classrooms they are registered to).

Table 14. Distribution of participants according to the institutional level

Institutional Level	Beginner	Elementary	Intermediate	Advanced	Very advanced	Total
N	13	16	25	21	8	83
Missing						2
N (UK)	9	6	9	4	3	31
Missing						1
N (USA)	4	10	16	17	5	52
Missing						1

As seen on Table 14, participants had a wide range of proficiency levels and thus, comprised a heterogenous sample in terms of proficiency. Note that participants were from approximately 20 different universities in three different countries in order

[19] Since the Canadian sample is quite small, samples from USA and Canada was combined as one sample from North America, and this one sample is referred as USA during the rest of this dissertation.

to reach a large sample. Therefore, there is possibly more variance than usual within the same institutional level.

The age of the participants had a wide range between 18 and 80 with a mean of 28.73. There were two unusual participants: one 80-year old and one 60-year old Turkish L2 learner. However, most of the participants were between 18 and 35 years old, showing the typical age students go to universities for various degrees.

Regarding the completed degree of education, the majority of the participants had completed a master's degree (N=29), followed by a bachelor's degree (N=26), a high school degree (N=22), and a doctoral degree (N=7). The subject of the completed degree of education varied from law to electronic engineering. However, the most common area of study was Middle East Studies (N=10), followed by Linguistics (N=7) and Turkish Studies (N=4).

The most common L1 among participants was English. Nevertheless, the sample of Turkish L2 learners were very heterogenous with 22 different L1s in total. There were 52 English, 6 Arabic, 3 Urdu, 2 Azerbaijani, 2 Finnish, 2 Greek, 2 Italian, and 2 Romanian L1 speakers. Furthermore, there was 1 L1 speaker of each of the following languages: Armenian, Bosnian, Dutch, Farsi, French, Georgian, German, Hebrew, Korean, Lithuanian, Mandarin, Norwegian, Russian, Uzbek. Finally, 3 participants identified themselves as heritage speakers of Turkish, and 7 participants identified themselves as bilingual speakers of Turkish. However, they did not comprise a large enough sample to be investigated separately, and they were not among the top scorers in the C-test.

The Turkish language background of participants (i.e., the age of learning Turkish, number of months of formal Turkish language learning) is given in Table 15

below. As seen, the sample of participants is very heterogenous in terms of all language background characteristics.

Table 15. Participants' Turkish background information

Variable	Mean	SD	Min	Max
Age of learning	23.89	7.97	13	52
Months of study	42.20	32.5	1.5	144
Months of residence in Turkey	8.85	13.32	0	72
Hours of study per week	5.97	12.78	0	80

6.3.2 SLA Researchers of Turkish

A total of 10 SLA researchers of Turkish participated in validation study 1. Of these researchers, 4 were from Turkey, and 6 were from the USA. Among them, there were 4 males and 6 females. Their age ranged between 28 and 45. All participating researchers, except one, read about C-tests before, and 3 of them previously used a C-test either in Turkish or English. Table 16 shows their academic ranks at their universities.

Table 16. Participants' academic ranks

Role	Graduate Student	RA/TA	Postdoctoral Researcher	Lecturer	Assistant Professor	Total
N	3	3	2	1	1	10

Of the SLA researchers, 5 participated in a follow-up interview. The interviews lasted between 14 and 28 minutes. The details of this group of interviewees are given in Table 17 below.

Table 17. SLA Researcher Interviewee Data

ID	Gender	Age	Country	Academic Rank	Research Interest	Interview Length
1	M	30	Turkey	Associate Professor	SLA	28

3	M	35	USA	Graduate Student	SLA	21
5	F	35	Turkey	Postdoctoral Researcher	SLA	16
6	F	25	USA	RA/TA	L2 Pedagogy	14
7	F	29	USA	RA/TA	Psycholinguistics	25

6.4 Instruments

The instruments used in this study for Turkish L2 learners involve two types of measures of Turkish proficiency, which are background questionnaire and Turkish C-test, as well as a follow-up feedback survey. The instruments for SLA researchers of Turkish involve a survey and interview questions.

6.4.1 Background Questionnaire for Turkish L2 learners

Turkish L2 learners completed an online background questionnaire for two reasons: (1) to find whether any construct-irrelevant variance resulting from factors such as first language or computer familiarization might influence test scores for the generalization inference; (2) to reveal any other indicators of L2 proficiency (i.e., months of residence in a Turkish-speaking country, length of Turkish L2 study) for the extrapolation inference.

The background questionnaire asked participants to respond to general demographic information, Turkish learning history and use, institutional level, and self-perceived proficiency in Turkish (revised from the test development stage, see section 5.5.2.1). The revisions were due to the changes in test administration method (online rather than paper-based) and including learners from UK as well. The revisions included the following: (1) a Likert-scale question about participants' level of comfort in using computers was included; (2) a new question about the country

where participants are currently studying / working in (UK or USA) was included; (3) the question about year in college was changed to completed year of education since UK and USA have different terms and length of college year; (4) a new question related to being a heritage and bilingual speaker of Turkish was included with short definitions of these types of speakers; (5) finally, a new question about learning difficulties related to reading and writing such as dyslexia was included since it could interfere with Turkish L2 learners' performance on the C-test (see Appendix 7).

6.4.2 The Turkish C-test

After completing the background questionnaire, L2 learners took the online 6-text Turkish C-test developed in Chapter 5 (see Appendix 8 for the 6-text Turkish C-test). Table 18 below shows the details regarding the level and content of each text. Note that texts are numbered the same way across the three empirical chapters to make comparison across studies easier later.

Table 18. Levels and content of the 6-text C-test

Text	ILR Level	Topic	Characteristics	Source
1	1	Locations	Very basic sentences with "there is/there are" structure Familiar words, cognates	Created based on commercial textbooks
3	1	A Danish person in Turkey	Simple sentences with present continuous Concrete words, some cohesive devices	Graded Turkish reader book (authentic texts)
4	1+	Description of a Turkish City	Simple sentences with relative clauses Informative social purpose	Adapted from an airline website
7[20]	2	Student success	Cause and effect relations Factual information	A health organization website

[20] Eliminated after analysis

9	2+	Relation between taste and smell	Conditionals and negations Topic specific vocabulary	Newspaper
12	3+	Relation between cultural venues and folk dances	Social and abstract topic Less-frequently used and topic-specific vocabulary Long and complex sentence structures	Academic journal in social sciences

Participants were provided with detailed written instructions including a practice item, information about the general format of the test, and recommended test taking strategies. Test takers were given a total of 30 minutes (5 minutes per text) to complete the test once they started it, which was reasonable given that learners spent an average of 42 minutes to complete the 11-text C-test in Chapter 5. Test takers could choose Turkish special characters (ç, ı, ğ, ö, ü, ş) from a text box when they were completing the gaps. They were warned not to use any external aids and to be careful about spelling since spelling counted. Texts were ordered according to their difficulty with Text 1 being the easiest and Text 12 being the most difficult in order to facilitate the familiarization of learners with the test format.

6.4.3 Feedback Survey for Turkish L2 learners

After completing the test, L2 learners were asked to complete an online feedback survey (see Appendix 9 for the feedback survey for L2 learners) to get learner input for the decision inference. The survey involved questions about participants' test taking experience and views about the Turkish C-test such as *Please select the level of difficulty for each C-test text.* It was the revised version of the post-test questionnaire used in the test development stage (see section 5.4.2.3). Revisions included the following: (1) a yes/no and short answer question about the user-friendliness and impact of the unsupervised internet testing; (2) a Likert-scale question about to what extent test takers think the Turkish C-test is a good and fair estimate of Turkish

language ability and an open-ended question about why they think so. The purpose of the feedback survey was to reveal test taking experience and any potential problems test takers might have encountered during test administration. Thus, it would be ensured that any experience or issue that might have influenced test takers' performance on the test was taken into consideration in data interpretation. At the end of the survey, test takers were asked whether they would like to receive an electronic Starbucks e-gift card for their participation.

6.4.4 SLA Researcher Survey

An online survey was administered to SLA researchers of Turkish in order to investigate whether the Turkish C-test is a useful tool for researchers by eliciting researchers' perception of the test (see Appendix 10 for the SLA researcher survey). This survey consisted of three sections: (1) multiple-choice and short answer questions about researchers' background (age, gender, country of location, L1, academic ranks); (2) yes/no and short answer questions about researchers' familiarity and experience with C-tests, as well as a Likert-scale question related to researchers' views about the usefulness and fairness of C-tests; (3) Likert-scale and open-ended questions related to researchers' views about the Turkish C-test (i.e., difficulty, usefulness, fairness, clarity). Researchers read an overview and example of C-tests in general as well as the instructions and texts of the Turkish C-test in sections 2 and 3. At the end of the survey, they were asked whether they would like to be reimbursed for their participation with Starbucks e-gift cards and whether they would like to participate in a follow-up online interview.

6.4.5 SLA Researcher Interview

Semi-structured interviews were conducted with 5 researchers who expressed an interest in a follow-up interview. All interviews, except one, were conducted via

Skype using Ecamm Call Recorder for Skype. One interview was conducted via Zoom because the interviewee preferred that option. The aim of the interview was to ask researchers to elaborate on their responses to the survey questions about their views on the C-test. The interviews lasted between 14 and 28 minutes. The questions were related to language assessment in SLA research, C-tests and the Turkish C-test. For example, there were questions such as *What do you think about language assessment in SLA research?* and *How often do you need to estimate your participants' proficiency levels?* (see Appendix 11 for the SLA researcher interview questions).

6.5 Data Collection Procedures

Participants were recruited through e-mail invitations. E-mails were sent to individual students and researchers, mailing lists (i.e., American Association of Teachers of Turkic languages), and Turkish language programs and instructors in different universities. A participation invitation with an overview of the research was also published in the January newsletter of the American Association of Teachers of Turkic Languages.

A list of US universities with Turkish language programs was reached through the enrolment survey for Turkic Language Courses in the US where 30 US post-secondary institutions were said to offer Turkish courses (Ergül, 2017). Also, another list of US universities with Turkish language instruction was obtained from the website of the Institute of Turkish Studies, and there were 41 universities on this list; however, it was not updated as the list in the enrolment survey. Unfortunately, no published list could be reached regarding UK universities with Turkish language instruction.

E-mail invitations were sent to 35 US and 9 UK universities as well as an academic language institution in UK. Of the US universities, 19 agreed to send e-mail invitations to students registered in their Turkish classes, 15 did not respond, and 1 rejected. Of the UK universities, 7 agreed to share the study link with the Turkish L2 learners registered in their programmes, 1 did not respond, and 1 said that they did not offer Turkish classes during that semester. Since the number of UK universities was smaller compared to US universities, a language institution (Yunus Emre Institute in London) which offered academic language courses aligned with the Common European Framework of Reference for Languages (CEFR) levels was also contacted, and it agreed to send e-mail invitations to their Turkish L2 learners.

The background questionnaire and surveys were created on Qualtrics, a secure software which is commonly used by researchers for data collection. Participants were informed about the goals of the study and their benefits from participating in this study in the participant information sheet and consent form on the first page on Qualtrics (see Appendix 12, 13, 14 for student, researcher, and interviewee information sheet and consent forms in turn).

Since Qualtrics did not support the structure of C-test (filling in gaps at every 2nd word in a text), the Turkish C-test was set on Learnclick (www.learnclick.com) which is a useful software to create various forms of language quizzes. There was one text per page, and the remaining time was shown on the screen. Learnclick recorded how much time test takers spent on the test as well as their responses to all gaps, and scores in each text. The average time participants spent on the 6-text C-test was found to be 20 minutes. Test takers were shown their total score and score percentage at the end of the test. Figure 14 shows a screenshot of how the test looks like on Learnclick.

Figure 14. Screenshot of Turkish C-test on Learnclick

6.6 Data Analysis Methods

The methods which were used to obtain evidence to support each inference and answer the related EQs are presented in this section. Note that analysis of the theoretical grounds inference is not included below (EQs1-2) since these are based on the literature review and the results relating to theoretical grounds are reported in the results of theoretical grounds section 6.7.1.

6.6.1 Analysis of the Scoring Inference

This section relates to the assumptions and EQs 3-7 under the scoring inference as well as methods used to answer these EQs. They are summarised in Table 19.

Table 19. Scoring Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods

3. Text selection and word deletion procedures are appropriate to cover a range of L2 learners in terms of Turkish general proficiency.	3. To what extent does the text selection and word deletion procedures produce a test that can cover a range of Turkish L2 learners?	3. Expert judgement, learners' perception, Rasch analysis (examinee separation indices)
4. Psychometric characteristics of texts are calculated, and the best functioning 5 texts are chosen for the final test version.	4. Which 5 texts discriminate between Turkish L2 learners of different proficiency levels most accurately and reliably?	4. Rasch analysis (item fit indices, item discrimination values, item difficulty measures, standard error estimates), Item-examinee maps
5. The C-test distributes test takers along a wide continuum of scores.	5. To what extent does the C-test elicit a wide range of scores?	5. Descriptive statistics of the total scores, graphical analyses, K-S Test
6. The scoring criteria are appropriate for the test.	6. Are the scoring criteria appropriate?	6. Answer key based on undeleted versions of the words and alternative answers, TUD
7. The scoring criteria are applied accurately and consistently.	7. Are the scoring criteria applied accurately and consistently?	7. Automatic scoring

EQ3 about the sufficiency of the text selection and word deletion procedures relates to the test development stage as well as learners' perception of the text difficulty collected in a survey in this chapter. The relevant details about the test development can be seen in Chapter 5, but it will be briefly summarised in this section. First, a total of 18 texts covering different levels of difficulty were chosen from authentic resources and a course book. They were assigned a level on the ILR reading skill scale by the researcher and a Turkish ILR level rating expert. Then, second-half deletion method was applied to these texts. After word deletion, the number of texts was reduced to 15 since the content in three of the texts was considered insufficient to reconstruct the words. 15 texts were piloted with Turkish native speakers. Based on their results, the number of texts was reduced to 11. The

final 11-text C-test was administered to Turkish L2 learners. The learners' test scores were analysed using Rasch analysis with FACETS (Linacre, 1989) (see section 4.5.2.1 for an explanation of Rasch Analysis). A two-facet (items and examinees) RSM (Andrich, 1978) was chosen to analyse the test scores considering all texts had the same 0 to 20 points scale despite their differing difficulty level. (see section 4.5.2.1 for explanations of IRT models). By using Rasch analysis, overall examinee separation indices (how many different levels of examinees the test is able to distinguish) were calculated. Also, descriptive statistics of learners' rating of text difficulty on a 5-point Likert scale were calculated.

Regarding EQ4, in order to come up with a final 5-text C-test test which discriminates between Turkish L2 learners of different proficiency levels (based on the initial institutional status) most reliably and accurately, the following psychometric characteristics of texts were calculated using Rasch analysis: item fit indices (whether the items fit the general pattern observed in the data), item discrimination values (how well the items are able to discriminate among test takers with different abilities), item difficulty measures and their standard error estimates. Furthermore, item-person maps were used to examine whether the test is well-targeted for test takers' ability.

In order to answer the EQ5 "*to what extent does the C-test elicit a wide range of scores"*, descriptive statistics of the C-test total scores (mean, median, minimum, maximum, range, standard deviation, skewness, kurtosis) were calculated using SPSS. Also, graphical analyses were consulted. Then, Kolmogorov-Smirnov (K-S) test of normality was conducted to: (1) to investigate whether the C-test distributes learners normally across a wide range of scores, (2) to see whether the sample of Turkish L2 learners are representative of a population with a wide range of Turkish L2 abilities.

Regarding EQ6 related to the appropriateness of the scoring criteria, the answer key was created based on the undeleted versions of the words and other alternative answers that emerged in the test development stage (see Chapter 5). In order to decide which alternative answers were acceptable, the researcher and a teacher of Turkish had discussions. When they could not reach an agreement over the acceptability of some answers, another teacher of Turkish and the Turkish National Corpus (Türkçe Ulusal Derlemi, TUD) were consulted. TUD is a web-based and large-scale corpus of Turkish language designed based on 50 million words (www.tnc.org.tr).

In order to answer EQ7 related to the application of scoring criteria accurately and consistently, the finalized answer key involving all acceptable answers was entered into Learnclick, and it automatically scored all student tests over 120 total scores (6 texts with 20 scores each) by using dichotomous scoring (1 or 0 depending on complete accuracy).

6.6.2 Analysis of the Generalization Inference

This section relates to the assumptions and EQs 8-11 under the generalization inference as presented in Table 20.

Table 20. Generalization Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods
8. The C-test texts are internally consistent, and they provide reliable estimates of test takers' L2 abilities.	8. To what extent does the C-test provide reliable estimates of test taker's L2 abilities?	8. Reliability analysis
9. The C-test functions consistently for Turkish L2 learners from both UK and USA.	9. Does the C-test produce consistent scores for both US and UK samples?	9. Reliability analysis and descriptive statistics both US and UK samples, independent samples t-test between two samples.

10. Texts are free of bias against any of the two groups.	10. Are texts free of bias towards UK and USA samples?	10. Descriptive statistics for C-test texts, DIF analysis
11. The sample of observations is large enough to control sampling error.	11. Is the sample of observations large enough to control for sampling error	11. Item-person maps, separation indices, literature review

First, in order to answer the EQ8 relating to the reliability of the C-test, Cronbach's alpha reliability coefficients were calculated for the initial 6-text C-test and the final 5-text C-test. Then, regarding EQ9 about the consistency of the C-test scores across USA and UK samples, reliability analysis of the final 5-text C-test was conducted for UK and USA samples separately. Furthermore, descriptive statistics (mean, SD, min, max) of the C-test scores for two groups were calculated, and an independent samples t-test was conducted between US and UK groups to ensure that the difference of mean scores between the two groups were not significant.

In relation to EQ 10 about the potential bias of any C-test texts towards the UK or USA sample, descriptive statistics of the C-test texts were calculated for each group. Then, Differential Item Functioning (DIF) analysis was conducted using WINSTEPS. The aim was to investigate whether any of the texts are biased against any of the two groups due to the differences in national curriculum systems (i.e., one text is unexpectedly more difficult for students from UK) (see section 4.5.2.3 for explanation of DIF). Texts that had DIF contrast (the difference in difficulty of an item between two groups) smaller than 0.50, and probability value bigger than .05 were identified as not having bias towards any of the two groups.

Finally, regarding EQ 11 about the sufficiency of the sample size to control for sampling error, item-person maps and separation indices derived from Rasch Analysis were also taken into consideration to interpret the sample size. Furthermore, literature was consulted to look at other validation studies conducted in LCTL.

6.6.3 Analysis of the Extrapolation Inference

This section explains the data analysis conducted for EQs 12-14 under the extrapolation inference as seen in Table 21.

Table 21. Extrapolation Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods
12. The C-test scores correlate with the variables of Turkish learning history and use derived from the background questionnaire.	12. Are there correlations between C-test scores and Turkish learning history as well as use?	12. Spearman's rho correlation coefficient
13. The C-test scores correlate with institutional level.	13. Are there correlations between C-test scores and institutional level?	13. Spearman's rho correlation coefficient
14. The C-test scores correlate with self-perceived proficiency in Turkish.	14. Are there correlations between C-test scores and self-perceived proficiency in Turkish?	14. Spearman's rho correlation coefficient

Correlational analyses were conducted to investigate the relation between the C-test scores and several other indicators of proficiency. First, in order to address the EQ 12 about the relation between C-test scores and variables of Turkish learning history and use (months of study, months spent in Turkey, age of learning, and hours of study per week), Spearman's rank-order correlation coefficient (Spearman's rho) was calculated between these variables and C-test scores because the variables did not have normality according to Kolmogorov-Smirnov test.

In order to answer EQ 13 about the relation between C-test scores and institutional level reported by students on the language background questionnaire, Spearman's rho was calculated due to the ordinal nature of institutional level. Finally, regarding EQ 14 about the relation between C-test scores and self-perceived

proficiency estimates on a 5-point Likert-scale, Spearman's rho was calculated again due to the ordinal nature of the Likert Scale.

6.6.4 Analysis of the Decision Inference

This section explains the data analysis conducted to partially investigate the decision inference and address EQ 15 shown in Table 22. Note that it was not feasible to fully investigate the decision inference and address EQ 16 since the Turkish C-test is yet to be used by several SLA researchers of Turkish in their research studies in order to explore whether it enables benchmarking, generalizability, interpretability, and replicability across SLA studies in Turkish.

Table 22. Decision Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods
15. The Turkish C-test scores reflect a certain degree of test takers' general language proficiency. They can be used to control for general proficiency levels of Turkish L2 learners in SLA studies.	15. What are the perceptions of the Turkish C-test stakeholders regarding the usefulness, difficulty, structure, and clarity of the Turkish C-test?	15. Thematic analysis of interviews and open-ended survey questions, descriptive statistics of Likert-scale survey questions
16. The Turkish C-test will enable benchmarking, interpretability, generalization, and replicability across SLA studies in Turkish for the proposed test use.	16. To what extent does the Turkish C-test enable benchmarking, interpretability, generalization, and replicability when it is used to control general language proficiency across SLA studies?	

Regarding EQ 15 about stakeholders' (researchers and learners) perceptions of the Turkish C-test, descriptive statistics of their responses to Likert-scale questions on the surveys were calculated. Furthermore, qualitative analysis was carried out on researcher interviews and open-ended questions of researcher and learner surveys. They were analysed by following the steps of thematic analysis recommended by

Braun and Clarke (2006). The details and the exact process of conducting thematic analysis was explained in section 4.5.3. To briefly summarise here, after transcribing all interview data and member-checking with interviewees, interview and survey data were read several times to be familiarized. Then, initial codes were identified by searching interesting aspects of data related to the EQ. Following this, themes and sub-themes were generated from initial codes by searching for repeated patterns within and across participants. These themes and sub-themes were reviewed to examine their relationship with each other and the broader EQ. Note that the analyses were conducted separately for both researcher and learner data.

6.7 Results

This section reports the results of the data analysis explained in the previous section 6.6 to answer the relevant evaluation questions. Results are presented under each specific inference that they relate to.

6.7.1 Results for the Theoretical Grounds Inference

This part presents the evidence for the theoretical grounds inference. Although no analysis was conducted for theoretical grounds and the theoretical evidence was based on the literature review, results relating to this inference are still reported by reflecting on the literature review and context to justify the connection of the construct of general language proficiency to the C-test format.

6.7.1.1 Components of General Language Proficiency (EQ1)

Several different models of language proficiency exist (i.e., Bachman, 1990; Hulstijn, 2015; see section 2.2.1). These models support the multicomponential nature of language proficiency; however, it is not certain what these components are and how they interact with each other in language use (Douglas, 2000; O'Sullivan & Weir, 2011; Purpura, 2008). Thus, the field of language testing still lacks a consensus

language proficiency model. All models of language proficiency, however, agree that the general components of language proficiency are grammar and lexis and they underlie more specific skills such as reading and writing. These general components were found to strongly associate with four language skills (Hulstijn, 2015). In the present research, general language proficiency is conceptualized as a unitary concept involving the common elements of grammar and lexis.

6.7.1.2 C-tests as a Quick Estimate of General Language Proficiency (EQ2)

C-tests involve the general elements (grammar and lexis) of language proficiency (see section 2.3.2). However, they cannot be reduced to a grammar or vocabulary test since they involve limited writing skills and require textual understanding to some extent due to their embedded and contextualized structure. A large number of studies have shown C-tests as measures of general language proficiency (i.e., Babaii & Ansary, 2001; Eckes & Grotjahn, 2006; Grotjahn, 2002; Norris, 2006, 2018; Harsch & Hartig, 2016).

C-tests were seen as integrative tests which requires combining multiple elements of linguistic knowledge. They were also found to associate with receptive and productive skills tests as well as discrete language skills such as grammar and vocabulary (see section 2.3.2.3 for correlations between C-tests and other language tests). However, since they are not based on a specific skill, domain, or task, the interpretation of their use should be limited to general estimations about language proficiency (Norris, 2018). Several factorial analyses showed that C-tests and other standardised language tests (i.e., TestDaF) load high on one single factor known as general language proficiency (i.e., Eckes & Grotjahn, 2006; Klein-Braley, 1994; Raatz, 1984). Sigott (2004) argued that which aspects of general language proficiency

C-test taps into depends on test taker proficiency as well as the level of text difficulty due to its fluid structure (see section 2.3.2.2 for details).

The Turkish C-test has been developed by following the C-test design principles summarised in 2.3.2.1 to tap into the C-test construct (see Chapter 5 for Turkish C-test development). Therefore, it is expected that the Turkish C-test reflects learners' general language proficiency in Turkish.

6.7.2 Results for the Scoring Inference

6.7.2.1 Text Selection and Word Deletion (EQ 3)

In order to answer the EQ 3, it is necessary to refer back to the test development stage in Chapter 5. Results of the initial investigation during test development are briefly summarised in this section. 5 texts were chosen out of the initial 11 texts based on their fit indices, discrimination and point-biserial values (see section 5.5.6.1). The reason to reduce the number of texts was to eliminate the misfit texts and make the test practical to fit within researchers' time constraints.

The 5-text C-test was able to distinguish between 5 different levels of learners (separation = 4.62; strata = 6.50 with .96 reliability (see section 4.5.2.2.1 for definition of these examinee statistics). All texts had infit and outfit statistics within the acceptable range of .5 and 1.5 (see section 4.5.2.2.2 for explanation of these item statistics). Furthermore, they all had high discrimination close to 1.0 and point-biserial values higher than 0.8, except the easiest text T1. The item-person map showed that there was a small group of high-level learners who was not covered even by the most difficult text. Therefore, a new high-level text was included in the C-test before the test was operationalized with a larger sample size in this validation study 1 (see section 5.6 for the inclusion of this text). This new 6-text C-test was piloted with Turkish native speakers (N=10) and they were able to complete all the texts, except

the new one T12, with at least 90% accuracy while the accuracy rate on the T12 was 82.22%.

A correspondence was also found between pre-estimated text difficulty and students' perception of text difficulty in the new 6-text C-test (1 as being very easy and 5 as being very difficult). Table 23 shows as the mean of self-perceived difficulty level gets higher from text 1 towards text 12. There is less standard deviation towards the most difficult

Table 23. Self-perceived difficulty of the texts by learners (N=81)

	Text 1 level	Text 3 level	Text 4 level	Text 7 level	Text 9 level	Text 12 level
Mean	2.20	2.58	3.47	3.81	4.31	4.78
Std. Error of Mean	.125	.120	.114	.121	.102	.050
Median	2	2	4	4	5	5
Mode	2	2	4	4	5	5
SD	1.12	1.08	1.03	1.09	.92	.45
Min	1	1	1	1	1	3
Max	5	5	5	5	5	5
Range	4	4	4	4	4	2

6.7.2.2 Psychometric Characteristics of C-test texts (EQ 4)

Rasch analyses were conducted by using 2-facet (examinee + item) RSM in order to come up with the best functioning 5 texts relating to the EQ 6. Table 24 shows the psychometric characteristics of the initial 6 texts[21].

Table 24. Key item quality statistics of 6 texts

[21] Text numbers are the same across all empirical chapters of this dissertation to make a better comparison afterwards. This is why they do not seem to be consecutively ordered here.

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.75	.70	1.13	1.36	.06	-.84
T3	.88	1.28	.65	.72	.05	-.23
T4	.92	1.37	.68	.64	.05	.43
T7	.90	1.19	.71	.81	.05	.27
T9	.84	.78	1.12	1.16	.06	.91
T12	.74	1.01	.90	.86	.07	2.14

As seen in Table 24, all texts submitted acceptable item fit indices and discrimination values between .50 and 1.50. They also had acceptable point-biserial values over .80 except the easiest Text 1 and the most difficult Text 12. Text 1 submitted similar results slightly below the acceptable point-serial and also the discrimination values in the test development stage as well. (see Table 9 in section 5.5.6.1). Note that T1 was the least authentic text created based on a dialogue in an elementary level Turkish course book.

Figure 15 shows the item-examinee map by putting item difficulty and examinee ability on the same scale.

```
+------------------------------+
|Measr|+Examinees|-Items|Scale|
|-----+----------+------+-----|
|   3 +          +      +(20) |
|     |          |      |     |
|     |          |      | 19  |
|     |          |      |     |
|     | *        |      |     |
|     |          |      | --- |
|     |          |      |     |
|     | *        | T12  | 18  |
|   2 + **       +      +     |
|     |          |      | --- |
|     | *        |      |     |
|     | **       |      | 17  |
|     | *        |      |     |
|     | *        |      | --- |
|     | *        |      |     |
|     | **       |      | 16  |
|   1 + ******   +      + --- |
|     | ***      | T9   | 15  |
|     | ***      |      |     |
|     | ***      |      | 14  |
|     | ****     |      | --- |
|     | ***      | T4   | 13  |
|     | ***      | T7   | --- |
|     | ****     |      | 12  |
*   0 * ****     *      * 11  *
|     | **       |      | 10  |
|     | ****     | T3   | --- |
|     | *        |      |  9  |
```

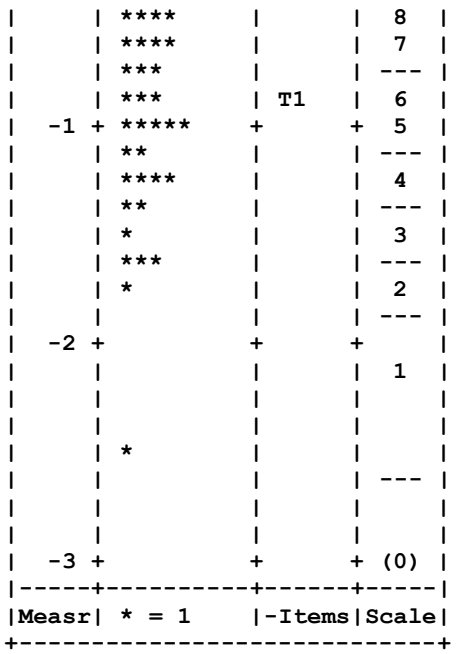


Figure 15. 6-text C-test item-examinee map

Higher level examinees and more difficult texts are displayed at the top of the figure while lower level examinees and less difficult texts are displayed at the bottom. 6-text C-test was able to spread examinees along a wide spectrum and distinguish at least 4 different levels of examinees (examinee separation= 4.39; strata= 6.18; separation reliability=.95). 91.01% of the total variance in C-test scores was also explained by the Rasch model.

As shown in Figure 15 and measure values in Table 24, the difficulty of each text was in line with the initial text ratings according to ILR levels (see section 6.4.2) except for Text 7 (measure=.27) being slightly easier than Text 4 (measure=.43). Upon analysing Text 7, it was found out that the word *daha* (meaning *more* in adverb form) was repeated five times on this text, and it was presented as an undeleted word three times. Crucially, one of these undeleted words was in the first intact sentence of the test, and one was at the beginning of the second sentence, which might have made it easier for examinees who did not necessarily know the word but saw it on the other sentences. As a matter of fact, the repetition of the word *daha* on this text and text 7

being easier than text 4 were also revealed while reporting the results at the test development stage (see section 5.5.6.1). Therefore, it was considered to change this word with the word *gayet* (meaning *pretty* in adverb form) which would suit the context and make the text more difficult (according to TUD, *daha* is 41 times more common than *gayet*). However, it was decided to keep the text as it is to stick to the originality of the text since *gayet* was not an exact synonym of *daha.* Furthermore, Text 7 had a difficulty measure of .03 which was quite different from other texts in the test development stage and therefore, there was no intention to make this text more difficult. However, as seen on Figure 15 above, Text 7 and Text 4 were found to be within the same grouping of difficulty. While having the same level of difficulty, Text 7 had lower point-biserial and discrimination values compared to Text 4. Therefore, Text 7 was eliminated in order to reduce the number of texts to 5 and make the test more practical. However, this does not mean Text 7 is problematic. It is just a redundant text not contributing to the overall difficulty of the test. Text 1 was not eliminated despite slightly falling under the acceptable point-biserial value because there was no other text to cover the lower level learners as shown in Figure 15.

After removing Text 7, Rasch analysis was conducted again with 5 texts. Variance slightly increased to 92.12% after the removal of the redundant Text 7. Separation indices slightly dropped, but still remained good by reliably distinguishing across 4 different ability levels of examinees (examinee separation= 4.01; strata= 5.68; separation reliability=.94). As seen in Figure 16, the newly added Text 12 was able to address the group of very high-level learners that was not covered by the previous 5-text C-test that was produced at the test development stage (see section 5.5.6.1).

```
+----------------------------+
|Measr|+Examinees|-Items|Scale|
|-----+----------+------+-----|
```

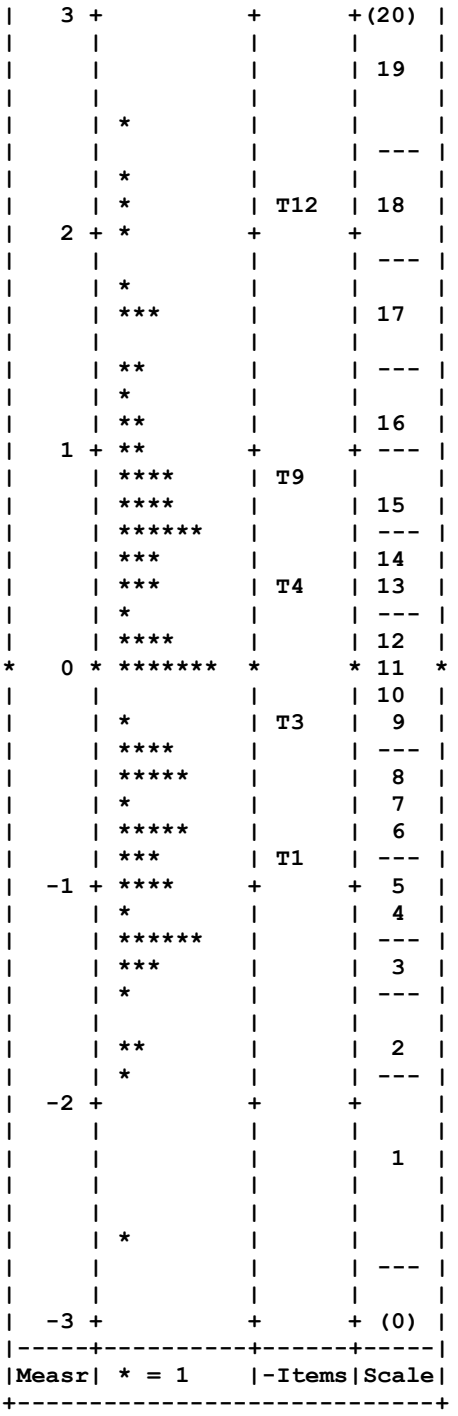


Figure 16. 5-text C-test Item-Examinee map

Table 25 shows the psychometric characteristics of the final 5-text C-test. All texts have infit and outfit statistics and discrimination values within the acceptable range of 0.5 and 1.5. They all also had point-biserial values over .80 except Text 1 and Text 12. Text 12 had a very different difficulty measure (2.14) compared to the second most difficult text T9 (.91) contributing to the higher limits of the test, which

might explain why its point-biserial values is slightly below the acceptable value. Overall, all texts had good item fit statistics, and no more redundant text with the same difficulty measure is left.

Table 25. Key item quality statistics of 5 texts

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.75	.84	.95	1.24	.06	-.84
T3	.87	1.27	.61	.70	.05	-.24
T4	.91	1.37	.63	.61	.05	.42
T9	.82	.74	1.08	1.23	.06	.91
T12	.74	1.06	.87	.86	.07	2.14

In addition to item quality statistics, individual examinee statistics were also examined (see section 4.5.2.2.1 for explanation of examinee statistics). Three outlier examinees were identified in terms of their performance (examinee infit and examinee outfit statistics higher than 2.0) after the Rasch analysis. One of these examinees was advanced (E15), one was beginner (E5), and the last one did not report her institutional level (E42). All the outliers were from the US sample. Looking at the feedback and background surveys completed by these examinees, interesting results were found, which might explain why they were identified as outliers. Firstly, E15 stated she had submitted an empty text (Text 9) due to a technical problem although she had filled the text out. Therefore, it is possible that her performance on the blank text was not consistent with her performance on other texts considering she was also an advanced level learner according to her institutional status. Second, E5 commented that she found all texts, except Text 3, extremely hard because she did not know enough Turkish, but text 3 contained more vocabulary that she was familiar with. Confirming what she commented, her score on Text 2 was exceptionally well compared to the ones on other texts. Note that Text 3 was not the easiest text according the item-examinee map on Figure 10. Lastly, E42 did really well on one of

the most difficult texts (17 out of 20 on Text 9) despite not doing so well on the first two easiest texts (15 out of 20 on Text 1 and Text3). Therefore, it is possible that she might not have tried equally hard on all texts. Her first language was Azerbaijani which belongs to the Turkic language family and has many lexical similarities to Turkish. Removing these outliers might be an option to slightly improve the results; however, it could also be interpreted as trying hard to fit the data to the model given the small sample size (N=85). Therefore, these seemingly outlier examinees were not removed (see Appendix 15 for Rasch analysis results when these three examinees are removed).

6.7.2.3 Statistics of the C-test total scores (EQ 5)

Descriptive statistics of the chosen 5-text C-test total scores regarding the EQ 5 about the sufficiency of the C-test to elicit a wide range of scores are provided in Table 26 below.

Table 26. Descriptive statistics of C-test total scores

		Total
N		85
K		100
Mean		42.35
Std. Error of Mean		2.29
95% confidence interval for mean	Lower bound	37.80
	Upper bound	46.91
Median		42
SD		21.12
Min		3
Max		87
Range		84
Variance		446.23
Skewness		.154
Std. Error of Skewness		.261
Kurtosis		-.979
Std. Error of Kurtosis		.517

As can be seen in Table 26, the 5-text C-test elicited scores ranging from 3 to 87 out of a total possible score of 100 from a group of 85 test takers. The slight

positive skewness (.154) indicates that more scores were grouped around the lower end of the distribution. Therefore, in order to assess the normality of C-test scores, the Kolmogorov-Smirnov test was used, and the normality could not be rejected, *D* (85) = .084, *p*= .200. Therefore, it was concluded that C-test scores were normally distributed spreading learners along a wide continuum. Figure 17 also shows the normal distribution of C-test scores with a curve close to a bell-shape.

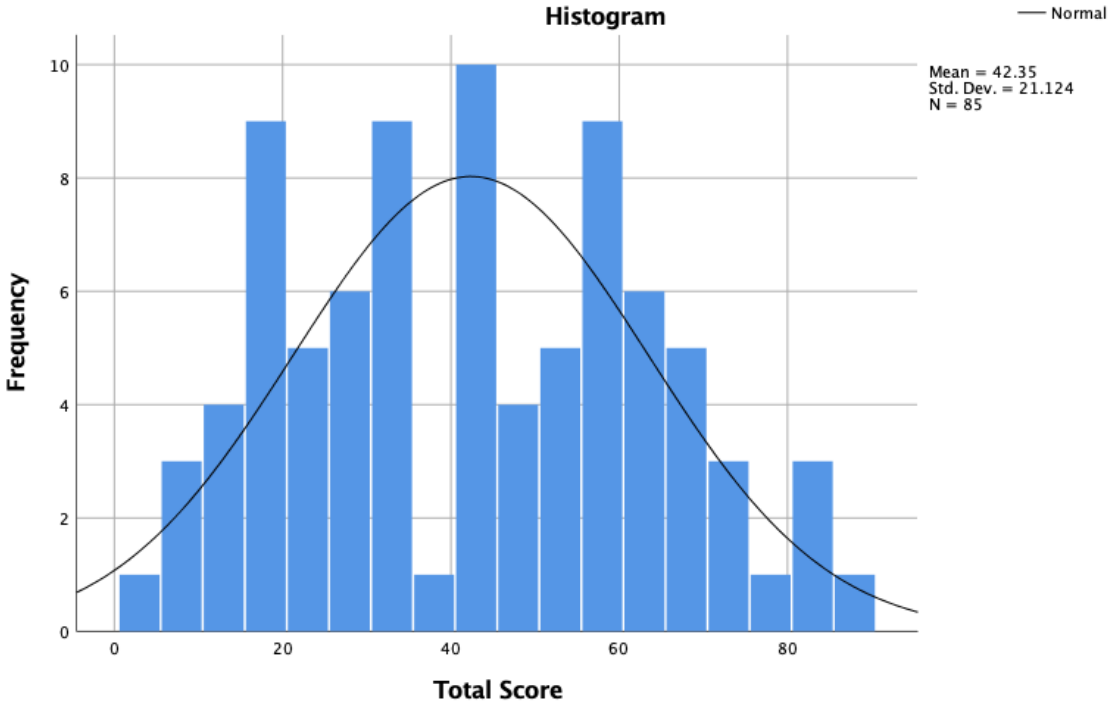


Figure 17. C-test score distribution

6.7.2.4 Appropriateness and Accuracy of Scoring Criteria (EQ 6 and EQ 7)

Regarding EQ 6 related to the appropriateness of the scoring criteria, the answer key proved to be appropriate to score the test taker answers. Looking at test taker responses to gaps closely, there were no more alternative answers in addition to the alternatives that came up beforehand. In relation to the EQ 7 about the application of scoring criteria, automatic scoring by Learnclick enabled accuracy, consistency, and objective results.

6.7.3 Results for the Generalization Inference

6.7.3.1 Reliability of the C-test (EQ 8)

A Cronbach's alpha of .92 was found for the 5-text C-test and a Cronbach's alpha of .94 for the 6-text C-test. As expected, reducing the number of texts slightly dropped the reliability estimate. However, in order to avoid overlapping texts in terms of difficulty and to produce a practical short-cut estimate of language proficiency, it was important to keep the number of texts to the minimum.

6.7.3.2 Consistency of scores across UK and US samples (EQ 9)

Reliability coefficients of the final 5-text C-test were also examined for each of the UK and US samples as seen on Table 27 below. Overall reliability coefficients remained consistently high for both groups despite reducing the number of texts.

Table 27. Reliability coefficients of the C-test

	UK (N=32)	USA (N=53)	All (N=85)
α for 6-text	.95	.93	.94
α for 5-text	.94	.91	.92

The descriptive statistics of the final 5-text C-test scores (out of 100 possible score) are shown in Table 28 for UK and USA samples separately again to investigate the consistency of the C-test across two groups.

Table 28. Descriptive Statistics of the C-test for UK and USA groups

		UK	USA
N		32	53
K		100	100
Mean		38.97	44.40
Std. Error of Mean		3.89	2.82
95% confidence interval for mean	Lower bound	31.04	38.74
	Upper bound	46.90	50.05
Median		33.50	49
SD		22	20.52
Min		8	3
Max		87	82
Range		79	79
Variance		483.77	421.128

Skewness	.695	-.177
Std. Error of Skewness	.414	.327
Kurtosis	-.355	-1.058
Std. Error of Kurtosis	.809	.644

Test takers from the USA sample had a higher mean (M=44.40, SE=2.82) compared to the UK sample (M=38.97, SE=3.89). Note that, US sample had a higher number of advanced and very advanced level learners (N=22) compared to the UK sample (N=7) although it would be ideal to recruit the same number of participants from each level in both samples. To investigate whether the difference between the mean scores of the UK and US samples was significant, an independent samples t-test was conducted. First, Kolmogorov-Smirnov test was used, and the normality could not be rejected in either the UK group, D (32) = .134, p = .152, or the USA group, D (53) = .104, p = .200. According to Levene's test, the variances of C-test scores in the two groups were equal (F =. 016, p= .899); therefore, no adjustments were made for equal variances. Finally, it was found that the mean difference between groups, -5.43, was not statistically significant, t (83) = -1.15, p =.253. As mentioned earlier, the insignificant mean difference might be due to that most of the beginner level participants in the total group was from the UK (see section 6.3.1 for information about participating L2 learners).

As seen on Table 28 above, minimum and maximum score levels were higher in the UK, and the UK sample had slightly more variance than the USA sample. Interestingly, both samples had the same range (R=79). Figures 18 and 19 show the distribution of C-test scores for the UK and USA samples separately. The data in the UK sample seem slightly positively skewed (.695), which might be due to the smaller sample size.

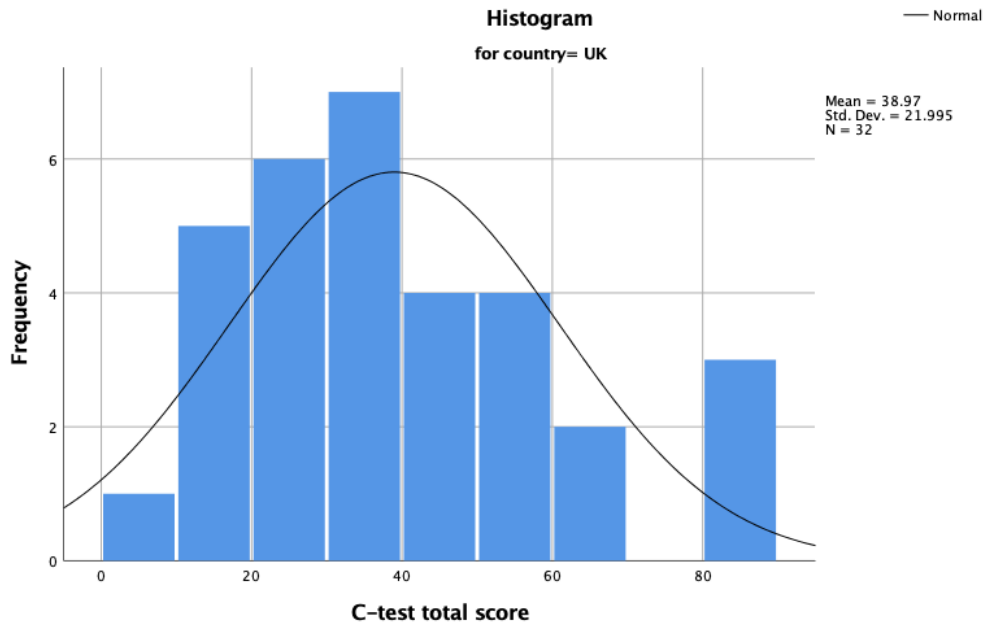


Figure 18. C-test score distribution for the UK sample

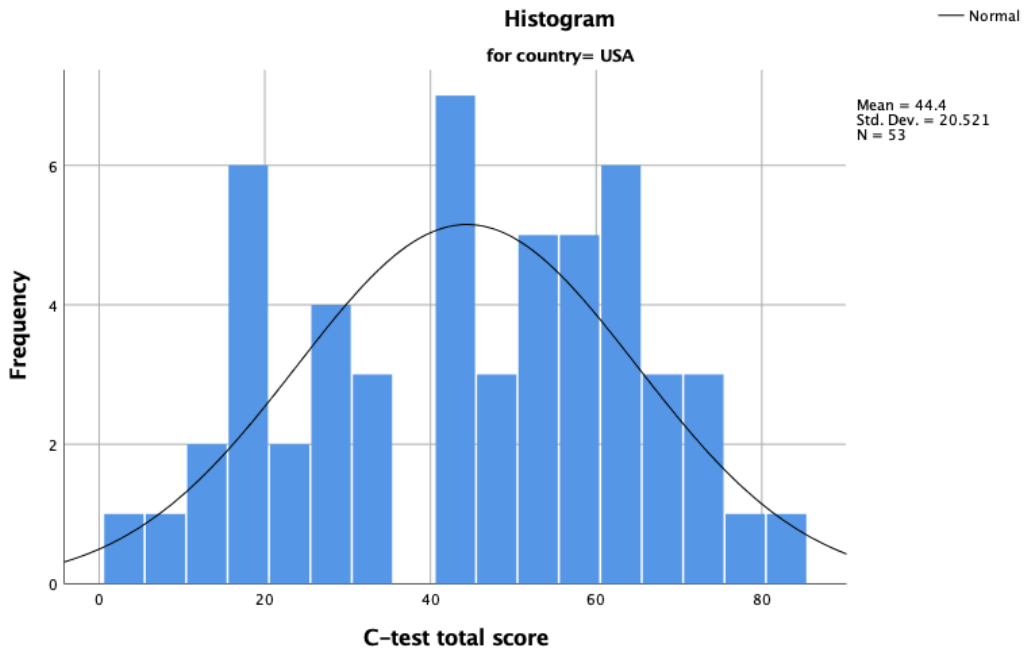


Figure 19. C-test score distribution for the USA sample

6.7.3.3 Investigating potential bias towards UK and US samples (EQ 10)

The next step was to investigate each C-test text for potential bias towards the UK or USA group relating to the EQ 10. Table 29 shows the descriptive statistics of each text for UK and USA groups separately.

Table 29. Descriptive statistics of texts for the USA (N=53) and UK (N=32) samples

	Text 1		Text 3		Text 4		Text 9		Text 12	
	US	UK	US	UK	US	UK	US	UK	US	UK
Mean	14.15	13.50	11.75	10.81	9.15	7.28	6.83	5.50	2.51	1.88
SD	4.53	4.32	4.78	5.18	5.62	5.80	5.57	5.73	2.97	2.98
Min	2	6	0	0	0	0	0	0	0	0
Max	20	20	19	20	17	18	20	10	11	11

Since the US sample had higher scores on each text compared to the UK sample, DIF analysis was conducted using WINSTEPS in order to investigate whether any of the texts were biased against the UK sample (see section 4.5.2.3 for explanation of DIF). There were two criteria to identify texts that had bias: (1) DIF contrast, which means the difference in the difficulty of an item between two groups, is bigger than or equal to 0.50, (2) the probability value, which means the chance of observing DIF contrast by chance, is smaller than or equal to .05. Figure 20 shows the difficulty of each text in logits for both groups. All texts seem to have a DIF contrast smaller than 0.50, with the highest DIF contrasts being in Text 3 and Text 12. The DIF measure table shows that DIF contrasts ranged between 0.00 and 0.25 and p values were bigger than .05 (see Appendix 16 for the DIF measure table). Overall, all texts were shown to function similarly for both UK and US groups without a noticeable bias.

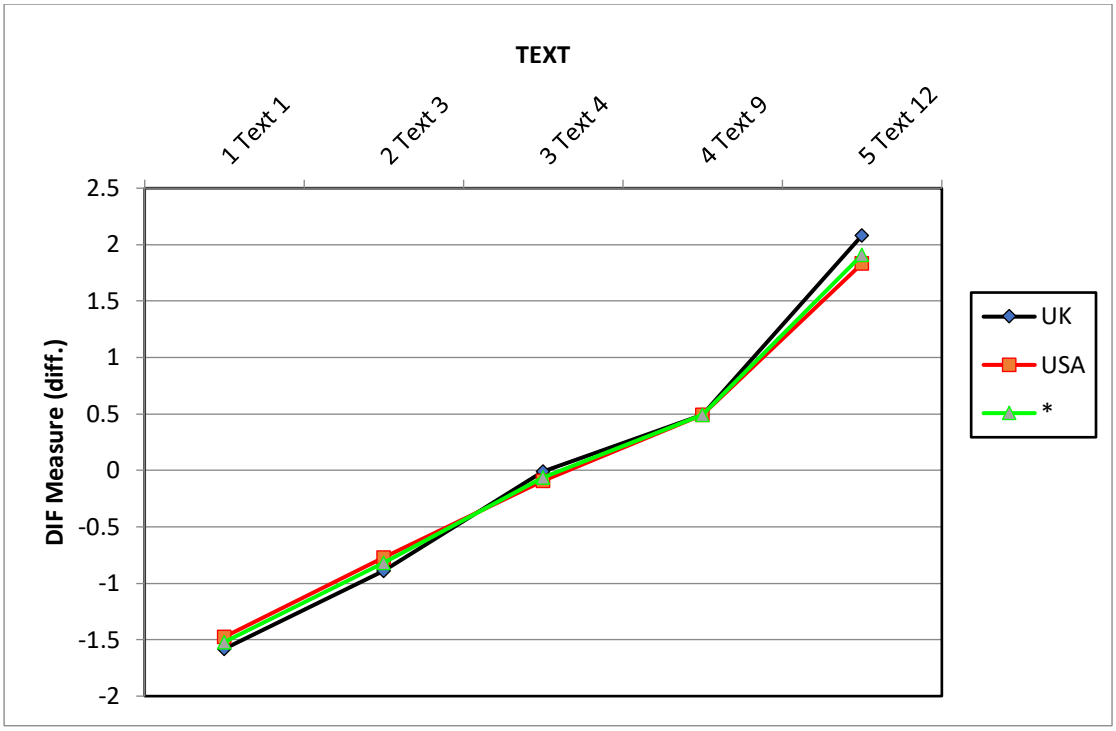


Figure 20. DIF plot[22]

6.7.3.4 Sufficiency of the sample size (EQ11)

No statistical analyses were conducted to investigate the sufficiency of the sample size to control for sampling error regarding the EQ 11. It is acknowledged that at least 10 observations are required per category for polytomous scores such as C-test texts which take a value between 0 and 20 (in this case, at least 210 participants). However, in less commonly taught languages such as Turkish, reaching such a big sample size is highly difficult. Nevertheless, the sample size of this validation study 1 (N=85) is similar to the sample sizes of other validation studies in LCTL which used IRT analysis as well as inferential statistics (i.e., Drackert, 2016; Son, 2018). Furthermore, IRT examinee separation indices (examinee separation= 4.01; strata= 5.68; separation reliability=.94) and item-examinee maps showed that there were at least 4 different ability levels involved in the sample (see Figure 16 in section 6.7.2.3)

[22] * indicates the expected baseline measure when there is no DIF contrast

6.7.4 Results for the Extrapolation Inference

This section reports the correlation coefficients between 5-text C-test scores and several other indicators of language proficiency involving language background variables, institutional level, and self-perceived proficiency in order to provide extrapolation evidence. Correlation coefficients were also calculated for 6-text C-test scores, and no discernible differences were found (see Appendix 17). Note that no correlational analyses were conducted between the C-test and another Turkish proficiency exam. This limitation is addressed in Study 2.

6.7.4.1 Correlations between C-test scores and language variables (EQ 12)

The language background variables correlated with the C-test scores involved months of studying Turkish, months spent in Turkey, hours of studying Turkish per week, and the age of learning Turkish. None of these variables were normally distributed according to the Kolmogorov-Smirnov test (D (79) = .171, p <.001 for age of learning; D (79) = .139, p =.001 for months of study; D (79) = .258, p <.001 for months spent in Turkey; D (79) = .337, p <.001 for hours of study per week). Therefore, Spearman's rho correlation coefficients were calculated, and results are provided in Table 30.

Table 30. Correlations between C-test scores and language variables

	Turkish C-Test Scores (5-Text)
months of study	.56
months in Turkey	.51
age of learning	-.46
hours of study per week	.24

Note: all correlations statistically significant, $p < .05$

The highest correlation of C-test scores was with months of studying Turkish, followed by months spent in Turkey, age of learning Turkish, and hours of studying Turkish per week. The correlation with the age of learning is negative due to the inverse nature of the relationship between the age of exposure to a language and the

proficiency in that language (Tremblay, 2011). These results suggest that the more time L2 learners spend on studying Turkish, residing in Turkey and the earlier they start learning Turkish, the higher their Turkish C-test scores are. Note that, correlations with 6-text Turkish C-test scores provide the exact same results except a slight difference regarding the hours of study per week (ρ=.25) (see Appendix 17).

6.7.4.2 Correlations between C-test scores and institutional level (EQ 13)

A large and significant correlation was found between institutional level and the 5-text C-test scores, ρ= .75, $p < .001$ (ρ=77 for the 6-text C-test). Note that, there is expected to be more heterogeneity than usual within the institutional levels since students were from almost 20 different universities in three different countries.

6.7.4.3 Correlations between C-test scores and self-perceived proficiency (EQ 14)

Table 31 shows that large correlations were found between C-test scores and self-perceived proficiency in the four main skill areas as well as overall self-perceived proficiency. These results suggest that C-test scores are related with both oral and written skills.

Table 31. Correlations between C-test scores and self-perceived proficiency

	Turkish C-Test Scores (5-Text)
Self-reading	.80
Self-writing	.82
Self-listening	.80
Self-speaking	.81
Self-overall	**.81**

Note: all correlations statistically significant, $p < .001$

6.7.5 Results for the Decision Inference

6.7.5.1 Perceptions of Stakeholders (EQ 15)

This section reports test stakeholders' (SLA researchers and Turkish L2 learners) perception of the Turkish C-test regarding its usefulness, difficulty, structure, and clarity.

6.7.5.1.1 SLA Researchers of Turkish

Initially, a series of 5-point Likert-scale questions elicited researchers' opinions regarding the: 1) clarity of the Turkish C-test example, 2) sufficiency of the Turkish C-test instructions, 3) appropriateness and fairness of the Turkish C-test to estimate general Turkish L2 ability, and 4) usefulness of the Turkish C-test in research studies to estimate general language proficiency levels (see Appendix 10 for SLA researcher survey). Of the 10 SLA researchers, 8 chose "strongly agree", and 2 chose "somewhat agree" for the clarity of the Turkish C-test example. Furthermore, 6 chose "strongly agree", and 4 chose "somewhat agree" about the sufficiency of the Turkish C-test instructions. Note that the sample size is small to excessively generalise these results.

Most of the participating researchers were positive about the Turkish C-test being a good and fair estimate of Turkish language ability and also using the Turkish C-test in their studies. Table 32 shows the Likert-scale statements and the percentage of researchers' agreement/disagreement with these statements.

Table 32. SLA researchers' perception of the Turkish C-test

	Strongly or somewhat agree	Neither agree nor disagree	Strongly or somewhat disagree

The Turkish C-test above is a good and fair estimate of Turkish language ability.	70%	10%	20%
The Turkish C-test above will be useful in my research studies to quickly estimate my participants' overall Turkish language proficiency levels.	80%	10%	10%

The open-ended survey questions, which asked the researchers to elaborate on their responses to Likert-scale survey questions, helped to elicit why researchers agreed or disagreed about the usefulness and appropriateness of the Turkish C-test. The two main interview questions "*What is your impression of the Turkish C-test?"* and *"Would you use the Turkish C-test as an estimate of overall language proficiency in your research studies? Why or why not?"* also corroborated the open-ended survey questions (see Appendix 11 for full interview questions). Based on the thematic analysis of these survey and interview responses, the following main themes were generated in relation to the test usefulness: 1) practicality, 2) measuring only some aspects of language, 3) test taker unfamiliarity (see section 4.5.3 in Methodology Chapter for steps of the thematic analysis).

Theme 1: Practicality

The first theme and associated subthemes and codes are presented in Figure 21.

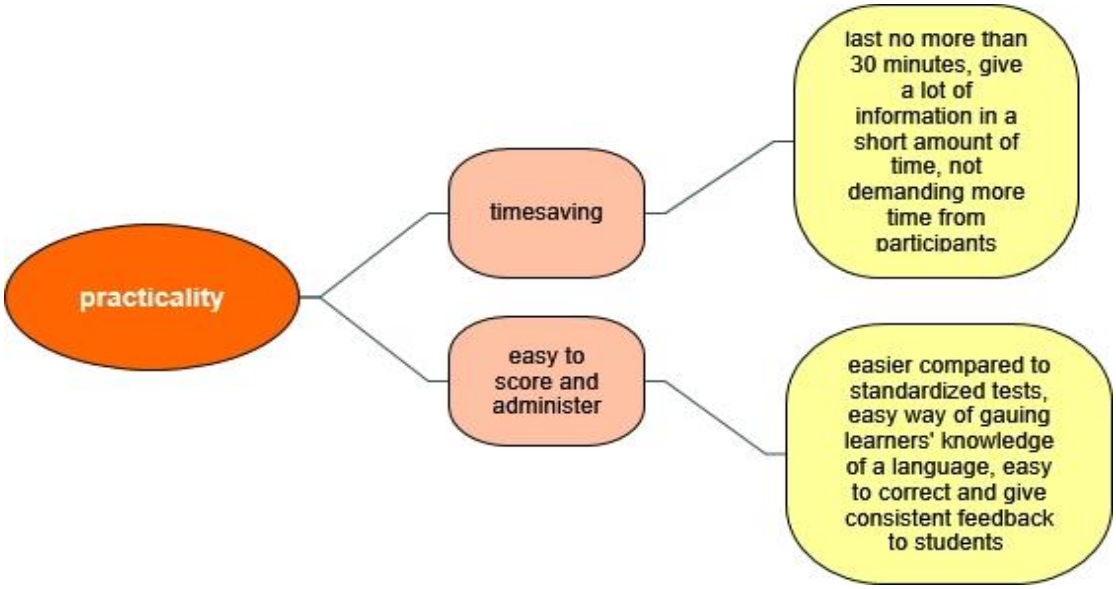


Figure 21. Theme 1 Practicality

Seven out of 10 participating researchers stated that the C-test is easy to administer and rate, and thus timesaving when they need a quick assessment of the proficiency. Words such as “quick”, “easy”, and “practical” kept recurring across interview and survey responses. As seen below, Researcher 1 stated that he would definitely use a C-test due to its practicality if it was valid.

> I guess if you have a valid C-test, it will be very timesaving because usually as far as I know C-test consists of a couple of texts or paragraphs do not take maybe more than 30 minutes. And rating or scoring seems to be easy, easier compared to other standardized tests. So, if there is a valid C-test, I would definitely use it.

Here, it is worth noting that validity relates to the interpretations and claims made based on the test scores (i.e., placing students into certain levels of language classrooms) rather than the test itself (Cronbach, 1971; Messick, 1989) (see section 2.4.5 for current views in validity). Thus, it is not possible to say that the Turkish C-test is valid itself. However, it can be used by SLA researchers in UK and USA when they want to control for language proficiency in their studies if they recruit a heterogenous sample of L2 learners in terms of language proficiency and the language proficiency may have an effect on the independent or dependent variable. As

explained in section 6.7.2.3, the Turkish C-test was able to distinguish across 4 different proficiency levels of Turkish L2 learners in the US and UK. However, the Turkish C-test may not be used when the language proficiency is a main variable since evidence related to the association of the Turkish C-test with golden standards of Turkish proficiency such as TYS was not collected in Study 1. This point was also made by Researcher 1 later during the interview.

> If the proficiency level is gonna be the main variable in my research, I may want to use another one maybe TYS. But if proficiency level will be just one of the subordinate or concomitant variables, then no need to look for another test. Instead, I would use a C-test.

Similarly, Researcher 5 mentioned that C-test gives a lot of information in a short amount of time. When asked whether she would use the Turkish C-test, she said that she would use it if it is found reliable. Note that the 5-text Turkish C-test was found to have a Cronbach's alpha of .92, which was in line with other studies investigating the reliability of the C-test where reliability estimates ranged between .75 and .96 (see Table 1 in section 2.3.2.3).

> For the past or the past 5 years, I am hearing more about C-tests as a tool of proficiency. I've seen researchers who found it reliable. And it is quick, practical, and it gives a lot of information. You need to be really automatic to find the possible words in that gap. It shows real use of language.
>
> I also searched for such a test actually. Whenever we do our research, we need to find something more you know professional. But of course, I have to see its reliability first. After you check its reliability, it would be better I think to see that. But it seems like a valid test to me.

Researcher 6, who was also teaching Turkish L2 classes, mentioned the effectiveness of the test as an easy way to conduct and give feedback to students.

> This test is very efficient in the sense that it is easy to conduct to evaluate the Turkish level of my students. It is relatively easy to correct and give consistent

> feedback while being not too difficult for the students thanks to the usage of easily guessable words.

Being fair to research participants also came up under discussions of practicality. Researcher 7 stated that it would be unfair to demand a lot of time from participants for comprehensive proficiency assessments since researchers have limited budget to reimburse participants.

> Like every academic would need a C-test, or some sort of a quick test, that is why I prefer to use a C-test or a cloze test rather than a proficiency test because we have a limited budget to give to participants. And I can't demand more time from them. C-tests are really easy to apply and less time-consuming, and maybe also to the point. I am really against using a very comprehensive proficiency test. There is no need. We don't have plenty of time, especially in psycholinguistic research. If I am gonna recruit participants, I shouldn't steal a lot of time from that person. That is unfair. So, that's why to make a fair study, I think we need a quick estimate of the proficiency. So, if you guys can develop like good ones, I am gonna definitely use them.

Overall, participants stated that they found C-tests timesaving, easy to administer and score, and fair to research participants in terms of their time; therefore, they would use the Turkish C-test if it is found to be statistically reliable and valid. When asked about in which contexts Turkish C-test would function better, researchers also considered that C-test would be most suitable to be used for research purposes, as a diagnostic test, or as the part of a placement test due to its practicality. Its potential use for teaching purposes also was mentioned. It is worth noting that developing a test "that fits it all", in other words a test that is valid for all purposes, is not possible. A test can be developed and validated for a specific test use and a particular group of test takers. Therefore, whether a test is valid depends on which purpose the test is being used for. As an instrument to control for adult Turkish L2 learners' language proficiency in SLA studies, the Turkish C-test was able to distinguish between four different proficiency levels in US and UK contexts and had a reliability estimate of .92.

Theme 2: Measuring only some aspects of language

The second theme was related to the format of the C-test as measuring only some aspects of language due to lacking an oral and writing component, and thus being insufficient to measure a learner's proficiency in a language as summarised in Figure 22. The interpretation of this limitation as good or bad depends on the researchers' goals and operationalization of proficiency as commented by participating researchers.

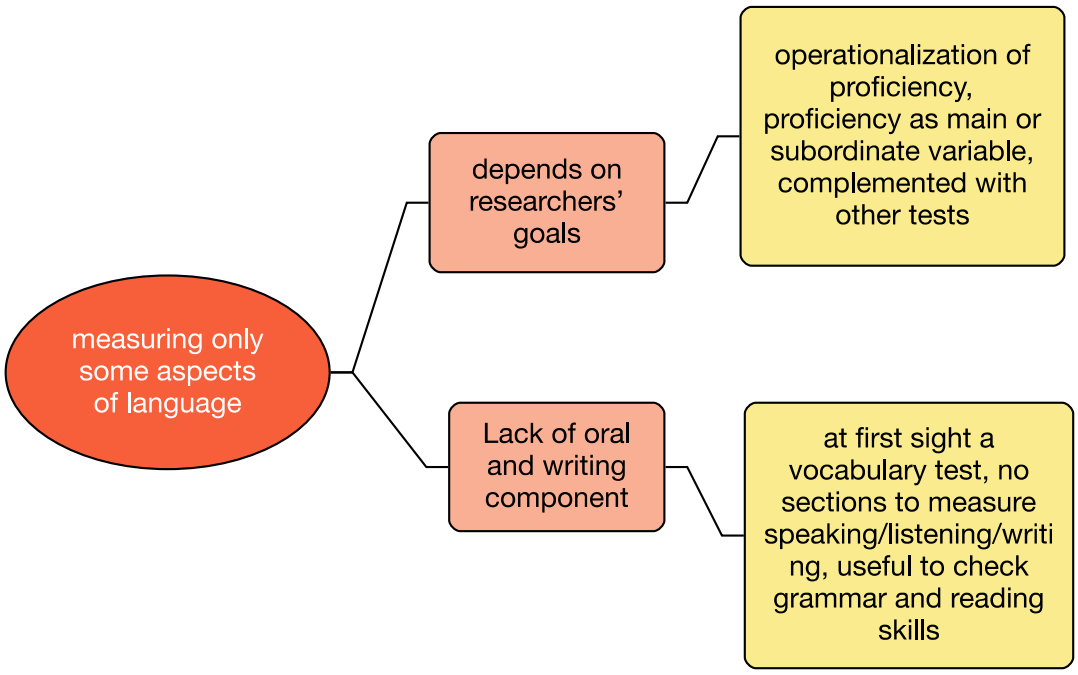


Figure 22. Theme 2 Lack of Oral and Writing Component

Researcher 1 stated his opinions about the limitations of what C-test can measure, and he added that due to these limitations he would look for another proficiency instrument if the proficiency was the main variable in his research.

> At first sight, it looks like a vocabulary test than a proficiency test. There are not sections addressing to measure speaking and listening skills. It looks like measuring only reading, vocabulary, and grammar. I am not sure how well it can check a learner's speaking, listening and even writing abilities. How you operationalize 'proficiency' appears to be a crucial step to see whether the c-test works well.

> If the proficiency level is gonna be the main variable in my research, I may want to use another one maybe TYS. But if proficiency level will be just one of the subordinate or concomitant variables, then no need to look for another test. Instead, I would use a C-test.

Similarly, Researcher 2 stated that C-test may miss measuring the crucial language skills that are necessary to be considered as successful language learners, and therefore, what the researcher aims to find out and how s/he operationalizes the language proficiency determines whether C-test can be useful in a given context.

> C-tests mostly rely on learners' knowledge of vocabulary, grammar, and reading skills. They do not test learners' listening, speaking and writing skills. For instance, intercultural communicative competence, interactional competence, as well as pragmatic competence are crucial skills that learners need to be considered as successful language users. But, C-tests may not gauge these skills due to its format. Therefore, it could be a quick and cheap (potentially free) estimate that can be used for certain studies. It really depends on the researchers' goals and how language proficiency is operationalized in the study

Several other researchers pointed out to the same limitation of C-tests as seeming more of a grammar, vocabulary, and reading test rather than a proficiency test.

> Researcher 4
>
> In order to assess a learner’s language proficiency accurately, oral and written and receptive and productive skills should be measured.
>
> Researcher 9
>
> I find C-tests quite mechanic and guided. They might work well if you want to check the students' reading and vocabulary skills; however, they are not useful for testing a learner's other major and minor skills in foreign language proficiency. By applying this kind of a test, you can hardly assess a student's speaking, listening, writing, etc. skills so it's difficult to have a fair say about a learner's proficiency in a language
>
> Researcher 10
>
> It seemed to me that C-tests provide a quick assessment of a learner's language ability in reading and writing skills rather than the overall proficiency. Reading and writing competence do not always automatically predict a learner's abilities in listening and speaking. This appears to be a potential weakness for C-tests.

In the SLA researcher survey, Researcher 3 said that C-test should be complemented with 1 or 2 other proficiency measures (i.e., listening comprehension test) since issues such as memory and word-recall might be involved

> I think C-tests are good, but they cannot adequately measure a person's proficiency in a language since issues such as memory, and word-recall are also relevant and crucial in successfully filling out information in a C-test. Hence, those tests should be complemented with 1 or 2 other proficiency measures.

However, later on during the interview he said his ideas somehow changed after he did the test himself. Interestingly, this researcher was the only researcher who also completed the Turkish C-test since he was also in the sample of native speakers who did the Turkish C-test. Other researchers were only required to review the test, but they did not have to complete it.

> My ideas somewhat changed compared to what I said earlier. After I completed your study, I realized actually C-test can be a good measure even just by itself even not complemented by other things because first I was thinking OK this can be related to memory and word recall but maybe not so much. Because if you are really proficient in a language, it is not going to be about word recall or like memory storage or something. It will be like whether you know the word whether you can meaningfully complete a sentence or not.

Overall, researchers seemed sceptical about what aspects of language proficiency C-tests can tap into. As previously mentioned in the literature review chapter (see section 2.3.2.3), C-tests in general lack face validity from the viewpoint of learners and teachers (Sigott, 2004). Lack of authenticity resulting from the test format was found among teachers' criticism of the C-test (Legenhausen, 1989). Furthermore, students felt that what the C-test measured was not clear and not necessarily within the construct of language proficiency such as imagination and inferencing (Huhta, 1996). To the best of my knowledge, no previous study has been conducted to

investigate researchers' viewpoint of the C-test. Nevertheless, this study showed that researchers' scepticism of the C-test aligned with previous studies investigating learners' and teachers' perceptions of the C-test. As some participating researchers of this study stated, the study purpose and the operationalization of language proficiency determines whether C-tests can be good and fair estimates of language proficiency.

Theme 3: Test-taker Unfamiliarity with Turkish C-test

The last theme was related to researchers' concerns about Turkish L2 learners' unfamiliarity with the Turkish C-test format and text content as expressed with words such as "shocked", "freak out", or "disappoint" describing learners' potential reactions. Figure 23 summarises this theme involving its subthemes and codes.

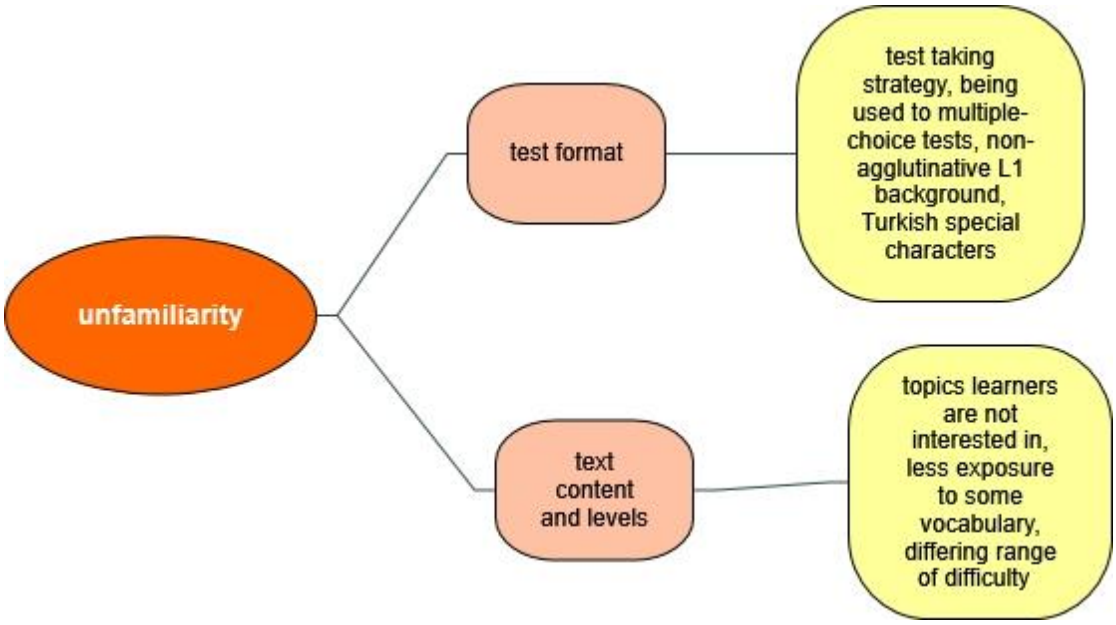


Figure 23. Theme 3 Test-taker unfamiliarity with Turkish C-test

Upon being asked whether the test is fair or difficult for learners, Researcher 1 said that his students in the department of English Language Teaching (ELT) would be shocked to see the C-test format since they got used to multiple choice tests before

coming to university as can be seen below. He added that students may find the test difficult although it is not that difficult due to their unfamiliarity with the test format. He also pointed out to the agglutinative structure of Turkish as a potential challenge for students (see section 3.3 for the definition and exemplification of agglutination).

> I mean it is also related to their test taking strategy. For example, we are going to have our first ELT students soon. In high schools, they are used to taking so many multiple-choice tests. If you give such a test to those students, probably they will be shocked. They are not used to this type of test. So, maybe it is the same for Turkish learners. If they are not familiar with this type of test, they might think it is a difficult test although it is not that difficult. That's one thing. The other thing, again it is also kind of related to that Turkish is an agglutinative language. Sometimes, they may find you know like deleted word, but not the whole one, so they may not be sure whether they got it correct or wrong after they are done with the test.

Similarly, Researcher 6 mentioned that the agglutinative structure of Turkish might be difficult in the C-test format for students coming from non-agglutinative languages since lots of suffixes are deleted from every second word in a text.

> Turkish is an agglutinative language and we use a lot of suffixes all the time. So, at times it might be difficult for students coming from like non-agglutinative languages because you just have to know all the suffixes out there, and the possibilities you have to create a reasonable word in that context, so it might be quite difficult for like new learners.

Turkish L2 learners' unfamiliarity with typing Turkish special characters (ç, ı, ğ, ö, ş, ü) also came up during the interview as can be seen below Researcher 3 pointed that the requirement to type Turkish special characters while also paying attention to what the correct word is might be cognitively too challenging for test takers, and they might freak out.

> You provided instructions like you have to use these special characters ç or ğ or ü or ö, and we don't have those letters in English alphabet. And you said something like you are going to lose points if you don't use those when you think those are the right letters. Of course, for a learner, they may freak out, they may be like where is ö where is ü, how do I do that. Because I work on Turkish linguistics, I use those characters quite frequently, so I am comfortable typing them or finding them. But for learners, it might be different. I remember while typing, you can just click on something and it

> gives you that letter, but it may still be too much. Should I worry about what the word is? Should I worry about trying to provide the correct spelling? It may be a little too much.

Researcher 7 was concerned that the differing range of difficulty across texts in one C-test might disappoint some learners if some texts are not at their level as seen below.

> I wanna have, as a researcher, a source where I can find a C-test for beginner level, a C-test for like intermediate level, because if I give a C-test to a beginner level student, but actually the C-test was designed for like advanced, then I don't wanna disappoint those learners at the onset of the study you know. I need their motivation, so I cannot make them feel disappointed at the very beginning of the study.

She was also sceptical that the content of the last text would be familiar and appropriate for even advanced test takers. However, the statistical findings showed that the last text (Text 12) was necessary to cover the advanced level test takers (see item-examinee map in section 6.7.2.3).

> One topic was really hard for me. It was the last one, I think. Because maybe I am not that interested in that topic in general in Turkish. Maybe, I don't read that much about that topic, so then that makes me less exposed to those vocabulary, or maybe that type of text style, I think the texts should be familiar to the reader, like the topic-wise. The topic of the last text gave me some hard time. If you give that C-test to even an advanced learner, what if they couldn't fill it in? For example, in my cloze test, I used the cut-off score 90%. There were 29 gaps, and I expected that 90% of the gaps should be filled in correctly. So, what if they cannot fill it in? Am I really gonna get advanced learners at the end?

Overall, researchers seemed somewhat concerned about the effect of the Turkish C-test on test takers if they were not familiar with the test format, text content, and typing Turkish special characters, or if the test involved texts that were beyond their level of proficiency. This finding is in line with the existing research on the face validity of C-tests in that teachers found the C-test too difficult for their students although actual student scores proved the contrary (Legenhausen, 1989;

Sigott, 2004). This discrepancy might be attributed to unfamiliarity with the test format and random selection of texts as earlier mentioned by Researcher 1 in this study. Note that the Turkish C-test was developed as a norm-referenced test to distribute different levels of learners along a continuum rather than a criterion-referenced test where there is an expectation for a high degree of accuracy. Furthermore, texts were chosen as neutral as possible without involving extensive technical and subject specific terms. However, advanced level texts such as Text 12 (ILR level 3+) unavoidably involve some subject-specific terms and cultural references.

6.7.5.1.2 Turkish L2 learners

Turkish L2 learners' (N=81) perception of the Turkish C-test was initially elicited through two 5-point Likert-scale survey questions: *"Select the level of difficulty for each text"* and *"The Turkish C-test above is a good and fair estimate of my Turkish language ability"* (1 as being strongly agree and 5 as being strongly disagree). Then, open-ended survey questions asked the learners to elaborate on their responses to Likert-scale questions (see Appendix 9 for learner survey).

There was a correspondence between pre-estimated text difficulty and students' perception of text difficulty (see section 6.7.2.1). More than half of the participating learners (54.3%) strongly or somewhat agreed that the Turkish C-test is a good and fair estimate of Turkish language ability as seen in Table 33.

Table 33. Learners' perception of the Turkish C-test (N=81)

	Strongly or somewhat agree	Neither agree nor disagree	Strongly or somewhat disagree
The Turkish C-test above is a good and fair estimate of Turkish language ability	54.3%	24.7%	21%

The open-ended survey questions helped to elicit why learners agreed or disagreed about the Turkish C-test being a good and fair estimate of Turkish language ability and why learners thought some texts are more difficult than others. Based on the thematic analysis of the open-ended survey responses, the following main themes were generated: 1) lack of relevant vocabulary and grammar, 2) measuring only some aspects of language (see section 4.5.3 in Methodology Chapter for steps of the thematic analysis). Note that the theme 'measuring only some aspects of language' was also found in researchers' interviews.

Theme 1: Lack of Relevant Vocabulary and Complex Grammar

The first theme was related to why learners found some texts more difficult than others as summarised in Figure 24. Forty-two participants (52%) reported that they lacked the relevant vocabulary, and twenty-six participants (32%) said that it was due to the increasing grammatical complexity.

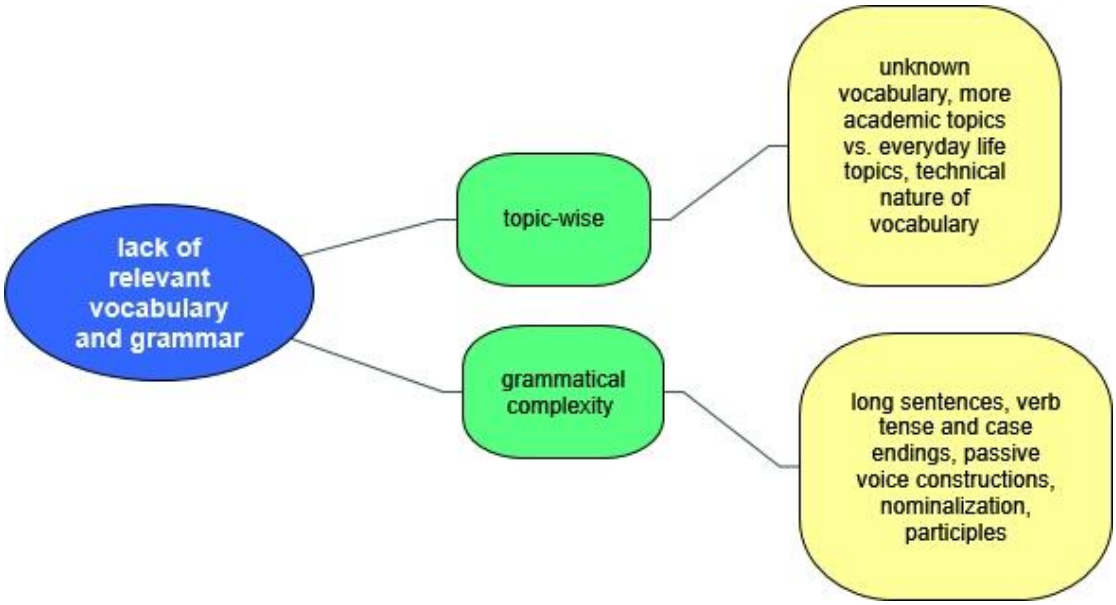


Figure 24. Theme 1 Lack of relevant vocabulary and complex grammar

As can be seen below, elementary level (based on self-reported institutional status) participant 45 said that he gave up after the third text due to his lack of vocabulary.

> I didn't have the vocabulary to answer the questions. Many choices just had one letter provided so I basically had to guess what the world could be, after third one I gave up because I just didn't know the words.

His test scores showed that he got 6 points from Text 1, 8 points from Text 3, 3 points from Text 4, and 0 points from the rest of the texts which corresponded with what he said. This is not surprising given that the first three texts which he completed were elementary level (ILR 1 and 1+) and then the texts started to get more difficult (ILR 2 and above). Remember that the Turkish C-test was designed as a norm-referenced test rather than a criterion-referenced test, thus it is expected that lower level learners would not be able to answer the texts that are above their levels.

Intermediate level Participant 2 and advanced level Participant 17 said that the latter texts were more difficult due to more complex grammar and vocabulary as can be seen below. Both of these participants had considerably lower scores on the last two texts compared to their scores on the previous text, which was aligned with what they reported on the survey[23].

Participant 2

> I was able to complete good portions of the earlier texts with confidence, based on my elementary or pre-intermediate vocabulary and grammar knowledge. Later the grammar constructions and vocabulary scaled up appropriately in difficulty.

Participant 17

> The latter texts were more difficult due to the more technical nature of the vocabulary used, as well as more complex grammatical structures.

Eight participants also mentioned that unfamiliarity with the text topic made some texts harder, because they did not know vocabulary related to that topic. Participant 51 mentioned disappointment with her more conversational knowledge.

[23] Participant 2 scores: text 1– 15, text 3 – 15, text 4 – 9, text 7– 9, text 9- 4, text 12 – 2. Participant 17 scores: text 1 – 17, text 3 – 16, text 4- 16, text 7- 16, text 9- 8, text 12– 4.

> Certain topics, obviously, require knowledge and good command of Turkish vocabulary in those topics; this is where I got stuck, and to be honest, disappointed with my knowledge, which is more day to day basic

She got lower points from the last two texts which had more academic topics compared to the previous texts: 2 points from Text 9 about the relation between smell and taste and 1 point from Text 12 relating to the cultural venues[24]. Topic unfamiliarity is a point also raised by one participating SLA researcher. As commented earlier, when texts got harder, it would be impossible to avoid some subject-specific terms and cultural references in texts.

Similarly, Participant 8 mentioned that she was clueless about the last text due to its topic and did not know vocabulary and suffixes in later [25].

> I had absolutely 0 understanding of what the last text said. It should be said, though, that reading is by far my worst skill!] The topics made some harder than others. For example, I have had conversations about student motivation before, because I am a teacher, but I have never talked about cultural continuity before - in Turkish or English, actually. I also don't know a lot of the suffixes and vocabulary used in the later texts because I was never taught them in a classroom.

Overall, it seems that participants found the later texts, in particular the last text, more difficult due to their unfamiliarity with the relevant vocabulary as well as topic and complex grammar structures. This finding is similar to the existing literature investigating the reasons of text difficulty from test takers' perspectives. Sumbling et al (2014) reported that test takers found the C-test difficult due to topic unfamiliarity, the lack of sufficient contextualization, and the limited test duration. Furthermore, as Legenhausen (1989) commented, test takers may consider the C-test as a criterion-referenced achievement test where they have an expectation of high accuracy while test takers construct C-test as a norm-referenced test having items of varying

[24] Participant 51 scores: text 1- 13, text 3- 10, text 4- 8, text 7 – 9, text 9 – 2, text 12 – 1.
[25] Participant 8 scores: text 1- 15, text 3- 10, text 4- 7, text 7- 7, text 9- 8, text 12- 1

difficulty. As previously mentioned, Turkish C-test is a norm-referenced test aiming to spread test takers along a continuum of proficiency. Therefore, there is no expectation for a high degree of accuracy from all test takers. To the contrary, it is expected that learners would not be able to complete the gaps in the texts above their level. Although test instructions stated that test takers may not fill in all the gaps, this may be made clearer in the instructions for future test administrations.

Theme 2: Measuring only some aspects of language

The second theme was related to participants' opinion about whether the Turkish C-test is a good and fair estimate of their Turkish language ability as summarised in Figure 25. There was variety in learners' views about what the C-test measures. However, the common view was C-tests measuring only some aspects of language since thirty-eight participants (47%) considered C-tests as measuring different skills such as reading and grammar rather than an overall proficiency. This perception might have resulted from the lack of an oral component, seeming lack of authenticity, and test takers' unfamiliarity with the test format.

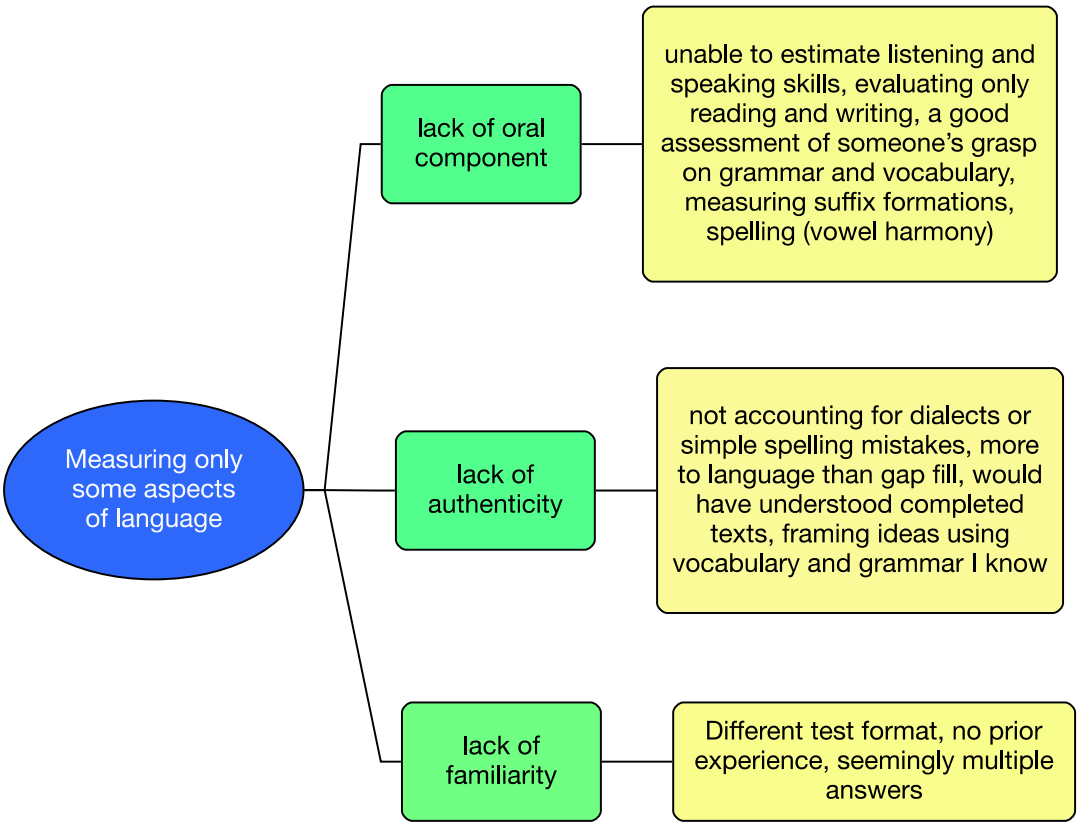


Figure 25. Theme 2 Measuring only some aspects of language

Eighteen participants (22%) mentioned the lack of an oral component, and thus C-test presenting an incomplete picture of language ability. As seen below, participant 8 stated that she is very good at communicating with Turkish speakers and does not think that Turkish C-test is very authentic due to its format.

> I learned almost entirely on-the-ground in Turkey. For this reason, I communicate quickly and [fairly] accurately with Turks, but I don't encounter a lot of the suffixes that are in these reading passages. Also, I rarely have to complete other people's words, and when necessary, I frame my own ideas using terms and grammar that I know. In other words, I don't think this is a very authentic assessment of true Turkish language usage. However, if we were assessing language ability for academic use (writing a dissertation in Turkish, for example), this may be more accurate.

Similarly, participant 52 implied the lack of authenticity due to gap filling format of the C-test.

> There is more to language than gap fill. I could answer the questions but did not always know what the sentence meant...This test doesn't necessarily capture someone's ability to speak or use the language.

Participant 9 also mentioned the lack of oral component and implied the lack of authenticity since scoring did not account for simple spelling mistakes or different dialects.

> I think that this does a fair job of assessing your reading and writing skills, but it doesn't cover speaking or listening at all. It also doesn't account for dialects or for simple spelling mistakes (which I don't believe should be counted as highly as not knowing the word at all).

Nine participants (11%) viewed the C-test as a reading test or reading and writing test as exemplified below.

> This test only evaluates the writing/reading part. Any language has other aspects such as speaking and listening. So, it's only assessing a part of Turkish language ability.

Participant 4 also emphasised that C-test may assess reading comprehension partially since the reading is integrated with writing while completing the gaps.

> Clearly this test is unable to estimate listening or speaking ability, which is a definite drawback. As for writing and reading, it does do a reasonable job of both prompting reading comprehension (in order to answer the questions) as well as writing (the actual act of answering). However, because the passages are read with blanks, it may not be a fair judge of actual reading comprehension, which is usually not impaired by the inability to produce written work.

Seventeen participants (21%) viewed the C-test as estimating grammar and vocabulary skills and thus not fully reflecting the language level as exemplified below.

> This is very good at estimating the grammar skills in Turkish and the formation of Turkish sentences, but some people are better at listening or speaking, so the Turkish C-test isn't a test that encompasses fully the level of language that you are at.

Participant 34 also mentioned the insufficiency of the C-test to distinguish among lower level learners, which aligned with the quantitative findings.

Nevertheless, the C-test does not aim to distinguish among learners in the same level, rather it aims to distinguish between learners in different levels as a norm-referenced test.

> It does test a student's vocabulary and ability to conjugate correctly based on context. However, I think it might do a poor job distinguishing between, say, someone that has never seen any Turkish and someone that has been learning Turkish for a few weeks because of the very all-or-nothing nature of the scoring. It also doesn't test for listening, speaking or writing (as in, how sentences are structured, what order words go in). However, I also know nothing about education and testing so this is all just guesses based on my own experience.

Participants' unfamiliarity with the test format might have contributed to their view of the C-test as an inauthentic and incomplete assessment of language proficiency. Statements such as "I've never seen a test like this", and "I have no idea what the C-test is" kept recurring among participants as exemplified below

> I have no idea what the C-Test is. I had not heard of it before taking the test. It seems like a test that doesn't require you to produce the language fully on your own might not be the best judge of your ability, though.

Furthermore, unfamiliarity with the C-test format led to some participants to be confused about whether there are multiple answers to gaps as seen in the excerpts below. Instructions of the C-test did not say that more than one answer was possible in some gaps to keep it concise; however, including this information could prevent some confusion on the side of the participants.

> Not sure if different answers were permissible or if there was only one right answer.

> In some cases, more than one answer was grammatically and semantically possible and I was unsure whether several options would be considered correct. As a result, I was hesitant about my choice, trying to guess what the expected variant would be.

Some participants also seemed sceptical about whether the system marks their alternative answers wrong as seen in the following excerpts. Participant 9 seemed

confident about her idea of multiple answers; however, her example is grammatically wrong since the word she wrote in parenthesis (mahallenin) is not written on the text. This is why the system marked her answer of “caddesinde” wrong. Regarding Participant 39, it is not obvious which blanks she is talking about. However, her answers were checked again, and there was not an answer that was true but marked wrong by the system.

> Participant 9
>
> I really think that some of the C-Test questions could have multiple answers. For example, when talking about the mahalle, the home could be "ana caddede" or "(mahallenin) ana caddesinde.
>
> Participant 39
>
> I think the test could ask a different format of questions since this format can cause confusion because the test taker may put a word that makes sense but get it wrong. Some of the blanks could be filled in in more than one way and I put a word that made sense but got it wrong according to the system.

In contrast with the common view about C-test measuring only some aspects of language, Participant 49 pointed out that one’s employment of vocabulary and suffixes in Turkish is a good measure of what they can achieve in speaking and writing.

> Endings are the foundation of Turkish and knowing enough vocabulary to navigate through those texts is a good measure to how well you could manage in both writing and speaking scenarios.

Along similar lines, Participants 3 and 59 said that the test reflected their self-assessment of Turkish language proficiency. This comment supported the quantitative findings which showed a significant correlation of .81 between C-test scores and overall self-perceived proficiency.

> Participant 3
>
> I have been struggling with learning Turkish. My wife, my colleagues, and my friends all speak English to a near native level which has unfortunately lead to

> my poor grasp of Turkish. Outside of "cafe" Turkish I am very weak, this test proved my self-assessment.
>
> Participant 59
>
> The test is a good way to assess a learner's competence in Turkish. It reinforced the conviction I had about my fluency in Turkish.

Overall, there was variety in learners' perception of what the C-test measured. This was in line with Sumbling et al's (2014) finding that there was no clear and consensus opinion among learners about whether the C-test is measuring general language proficiency. This lack of consensus among learners might result from two factors. First, as Sigott (2004) suggested, C-test has a 'fluid' structure and which aspects of L2 proficiency it taps into depends on learners' proficiency (see section 2.3.2.2). Second, most of the participating learners were unfamiliar with the unique format of the C-test, which might result in their views of the C-test as an incomplete and inauthentic task. Stakeholders' view of the C-test as an inauthentic task was also found in other C-test studies (i.e., Legenhausen, 1989; McBeath, 1989; Sumbling et al, 2014).

6.8 Discussion

This section discusses and summarises the results in order to answer to what extent each assumption was accepted or rejected under the interpretive argument. The results involve both a priori validity evidence presented in Chapter 5 (i.e., test development decisions made before data collection) and a posteriori validity evidence presented in the current Chapter 6 (i.e., correlation of test scores with other indicators of proficiency). They are discussed under each relevant inference.

6.8.1 Theoretical Grounds

The Turkish C-test was designed as independent from a specific language domain (i.e., business language use) to be used for research purposes in SLA, in other words,

it was not tied to a specific language curriculum or program. Thus, the texts involved both every-day (i.e., description of a city) and relatively academic topics (i.e., relation between smell and taste) depending on the text difficulty. Since there was not a specific domain to define, the first inference 'theoretical grounds" (Son, 2018) was evaluated considering theoretical justifications based on the literature.

First, the construct of general language proficiency was defined as a 'unitary' concept (Oller, 1971) involving but not limited to the general factors of language proficiency, namely grammar and lexis (Carroll, 1993). This was based on the evidence that all L2 proficiency models agreed on these general (core) elements of language proficiency although there was not a consensus L2 proficiency model (see section 2.2.1 for details on L2 proficiency models). Essentially, C-tests are a core test (Hulstijn, 2012, 2015) that requires the knowledge of grammar and lexis in an embedded context. However, they cannot be reduced to a grammar or vocabulary test since they also require limited writing skills and textual understanding. Furthermore, they also involve metacognitive strategies in addition to language knowledge and skills (Harsch & Hartig, 2016). Sigott (2004) found that whether learners use micro-level and macro-level cues while completing the gaps depends on their proficiency due to the 'fluid' structure of C-tests. For example, while some test takers use the whole passage at the text level while completing a gap, more proficient learners can operate with more limited context (i.e. word level, sentence level) to complete the same gap.

C-tests were found to strongly correlate with language skills tests (listening, reading, writing, speaking) as well as discrete-point grammar and vocabulary tests (i.e., Eckes, 2014; Eckes & Grotjahn, 2006; Harsch & Hartig, 2016; Sigott, 2004). They were also shown to load on the same single factor, namely general language

proficiency, as other language skills tests through confirmatory factor analyses (i.e., Eckes & Grotjahn, 2006; Klein-Braley, 1994; Klein-Braley & Raatz, 1984). Therefore, many studies used C-tests to provide a general estimate of language proficiency both in SLA and educational assessment (i.e., Dörnyei & Katona, 1992; Eckes & Grotjahn, 2006; Eckes, 2014; Lee-Ellis, 2009; Norris, 2006, 2018). These findings formed the theoretical grounds for the development and suggested uses of the Turkish C-test.

6.8.2 Scoring

The scoring inference links test takers' observed performance on the Turkish C-test to their C-test scores which reflect their general language proficiency in Turkish. It is based on six assumptions: (1) Text selection and word deletion procedures are appropriate to cover a range of L2 learners in terms of Turkish general proficiency; (2) Psychometric characteristics of texts are calculated, and the best functioning 5 texts are chosen for the final test version; (3) The C-test distributes test takers along a wide continuum of scores; (4) The scoring criteria are appropriate for the test; (5) The scoring criteria are applied accurately and consistently. The evidence to support these assumptions came from the descriptive statistics and item parameters of the test scores.

First of all, the appropriateness of word deletion and text selection was evaluated by investigating examinee separation value and reliability as well as analysing learners' perception of text difficulty. The initial 5-text C-test was shown to cover approximately 5 different ability levels during the test development stage in Chapter 5 (N=37). Linacre (2007b) suggested that if the examinee separation is lower than 2 and reliability is smaller than .80, the test may not be sensitive enough to discriminate between high and low levels. The values of examinee separation and

reliability was highly above these suggested threshold values. It proved that the second half deletion method and text selection using ILR guidelines was able to create a Turkish C-test that can address different levels of learners. Furthermore, the pre-estimated text difficulty aligned with students' perception of text difficulty.

However, item-examinee map showed that there was a group of high-level learners not covered by the most difficult text. It is worth noting that this type of ceiling effect was commonly observed in other studies investigating the discriminative power of C-tests as well as cloze tests (i.e., Grotjahn, 1987; Klein-Braley, 1985; Oller & Conrad, 1971; Tremblay, 2011; Son, 2018). Nevertheless, a new text at a higher level (ILR 3+) was included in the final version of the test. The reason why a higher-level text at ILR 3+ was not included in the initial version is that 90% touchstone level of native speaker accuracy was seeked after in all the texts during the test development stage. However, considering the group of high-level learners not covered by the most difficult text, 80% accuracy, as suggested by McKay and Abedin (2018), was taken as criterion in this validation study.

Regarding the second assumption about choosing the best functioning texts, the new 6-text C-test was administered to a larger sample size (N=85) in the validation study described in this chapter. Item-examinee map showed that there were texts in five different groupings of difficulty. However, one of the texts was redundant as it did not contribute to the difficulty of the test and had slightly lower discrimination and point-biserial values compared to another text of the same difficulty. Therefore, it was eliminated. The final 5 text C-test was able to distinguish across 4 different ability levels of examinees with a high reliability. Examinee separation was slightly lower than the initial administration during test development although a new text with a higher level of difficulty was added. This might be

attributed to the larger sample size. Overall, the final 5-text C-test had acceptable fit values satisfying the second assumption. All texts also had acceptable point-biserial values over .80 except the easiest (Text 1) and the most difficult texts (Text 12). The newly added Text 12 had a very different difficulty measure compared to the second most difficult text T9 contributing to the higher limits of the test, which might explain why its point-biserial values is slightly below the acceptable value. Analysing learner surveys, it was also found that some learners found the topic of the text 12 technical and inaccessible. However, when the texts get harder above ILR level 3, it would be impossible to avoid some subject-specific terms and cultural references in texts. Furthermore, thanks to the text 12, there were no more any high-level examinees not covered by the test. Regarding Text 1, it was the easiest text with the lowest difficulty measure, and the only inauthentic text created based on a textbook. However, it was essential to keep this text since it targeted the lower level learners.

Descriptive statistics of the C-test total scores provided the evidence for the third assumption about the power of the C-test to distribute learners. The C-test total scores were found to be normally distributed having a wide range. This finding contradicted some SLA researchers' concerns about the test being beyond the level of learners. As commented earlier, the C-test was developed as a norm-referenced test to distribute learners along a range of scores rather than a criterion-referenced achievement test. Findings showed that the test was fit for that purpose.

Regarding the assumption about the appropriateness of the scoring criteria, the answer key, which consists of the undeleted versions of the words and other alternative answers emerged during piloting, proved to be appropriate to score the test taker answers. Looking at test taker responses to gaps closely, there were no more

alternative answers. Furthermore, automatic scoring by Learnclick (the online test administration platform) enabled accuracy and consistency.

6.8.3 Generalization

The generalization inference links learners' observed C-test scores to their expected scores across C-test texts. It is based on four assumptions: (1) The C-test texts are internally consistent, and they provide reliable estimates of test takers' L2 abilities; (2) The C-test functions consistently for Turkish L2 learners from both UK and USA; (3) Texts are free of bias against any of the two groups; (4) The sample of observations is large enough to control sampling error. The evidence to back these assumptions came from reliability analyses, inferential statistics, and DIF analyses.

Regarding the first assumption, the internal consistency of the Turkish C-test texts was high aligning with other studies measuring the reliability of C-tests (i.e., Eckes, 2014; Jafarpur, 2002). As Roever (2018) noted, high reliability is one of the strengths of C-tests, and it was shown to exceed the reliability of well-established high-stakes exams in most cases. The high reliability of C-tests results from the consistency in their design, in that, the same text deletion and selection principles was applied to each text (Roever, 2018). The reliability of the Turkish C-test was also consistent across both UK and US samples. Independent samples t-test showed that there was no statistically significant difference in the test scores of these two groups, which provided evidence for the second assumption. Furthermore, none of the texts had any bias against US and UK samples as shown by the results of the DIF analysis.

Finally, regarding the assumption about the sufficiency of the sample size, it is acknowledged that at least 210 participants would be ideal considering 10 observations are required per category for polytomous scores of C-test texts which take a value between 0 and 20. However, in less commonly taught languages such as

Turkish, reaching such a big sample size is highly difficult since there are not enough number of learners and the sample size of this study is similar to other validation studies conducted in LCTL (i.e., Drackert, 2016; Son, 2018). A solution for this limitation would be replicating this study with a larger sample size by conducting the study over a longer period of time.

6.8.4 Extrapolation

The extrapolation inference links learners' scores on the C-test to their Turkish level on other indicators of general language proficiency. It is based on three assumptions: (1) The C-test scores correlate with the variables of Turkish learning history and use derived from the background questionnaire; (2) The C-test scores correlate with institutional level; (3) The C-test scores correlate with self-perceived proficiency in Turkish. The backing to support these assumptions came from correlational analyses.

Addressing the first assumption, there were significant and moderate correlations between the C-test scores and language learning variables (length of study, time spent in Turkey, age of learning, weekly hours of studying Turkish) as detailed in section 6.7.4.1. The correlations were higher than the ones found in Drackert (2016) when EIT scores were correlated with the same language learning variables. Note that while both EIT and C-test belong to the group of reduced redundancy tests, EIT is in the oral format and C-test is in the written format. Thus, C-test requires literacy skills which are gained through formal study of the language, and these skills may not be necessary for EIT. The highest correlation of the C-test scores was with the length of study, which is focused on explicit learning. Regarding the second assumption, there was a moderately strong and significant correlation between test scores and institutional level. However, it was not as high as it was in the test development stage (see section 5.5.6.2). This might be attributed to that there was

possibly more variance than usual within the same institutional level in this validation study. Participants were recruited from approximately 20 different universities in two different countries in order to reach a larger sample. Therefore, the results should be interpreted carefully considering the possible heterogeneity within the same institutional levels.

In relation to the third assumption, there were strong and significant correlations between C-test scores and self-perceived proficiency in listening, reading, writing, and speaking as well as self-perceived overall proficiency. All correlations were found to be close to each other suggesting that C-test scores are related with both oral and written skills. They were slightly lower than the correlations found in the test development stage, except the correlation with self-perceived proficiency in writing (see section 5.5.6.2), but higher than the correlations with self-assessment found in other languages in Norris (2018) (see section 2.3.2.3)

6.8.5 Decision

The decision inference links Turkish C-test scores to the intended use of the Turkish C-test through the input of test stakeholders and evidence collected for previous inferences. It is based on two assumptions: (1) The Turkish C-test scores reflect a certain degree of test takers' general language proficiency and can be used to control for general proficiency levels of Turkish L2 learners in SLA studies; (2) The Turkish C-test will enable benchmarking, interpretability, generalization, and replicability across SLA studies in Turkish for the proposed test use.

The evidence collected for the previous inferences provided partial support for the first assumption. Also, although SLA researchers found the Turkish C-test practical to use in their studies, they were sceptical about which aspects of language proficiency C-tests can tap into. Despite their strengths in psychometric

characteristics, the biggest weakness of C-tests lies in stakeholders' unfamiliarity with what they can measure and be used for (Roever, 2018). C-tests are not commonly used in mainstream testing, except in German university settings, in spite of the extensive research about its uses (Sumbling et al, 2014). Therefore, unfamiliarity with the test and scepticism towards its uses remains a problem contributing to its low face validity. To address this issue, it would be useful to allocate time to familiarize test users with the test uses and construct before administering the test. This can be done by providing test users with a practice C-test consisting of several texts. If time and location conditions allow, the researcher can go through these examples with test users explaining what kind of test taking strategies can be used

Regarding the second assumption, no evidence was collected. In order to fully explore whether the Turkish C-test can enable benchmarking, interpretability, generalization, and replicability across SLA studies, it is yet to be used by several SLA researchers as a proficiency tool in their actual research studies. However, given the time limitations, this was not feasible within the scope of this study.

Overall, the Validation Study 1 provided evidence for the use of the Turkish C-test as an instrument in SLA research studies to control language proficiency. Since the test use is not directly focused on language proficiency, such as using the test to select participants into a study based on their language proficiency levels, evidence regarding the correlation of the Turkish C-test with another test of language proficiency was not required. Nevertheless, some researchers were sceptical about the sufficiency of the Turkish C-test since it did not have an oral component. Addressing these concerns and using the test for determining language proficiency levels directly, the Validation Study 2 reported in the next chapter sheds light on the association of the Turkish C-test with speaking, listening, reading, and writing skills by investigating

its relation with a standardised Turkish proficiency test involving main language skills.

CHAPTER 7: VALIDATION STUDY 2

7.1 Introduction

Due to recent global developments such as immigration and scholarship programs, there has been an increasing number of international students in Turkish-medium universities (see Chapter 3 for more information). Thus, one of the improvements to facilitate the admission and enrolment of international students into these universities was the development of the TYS as an assessment of Turkish language proficiency. Students can achieve one of the following levels on the TYS depending on their total score: B2 (55-70 points), C1 (71-88 points), or C2 (89-100 points). Below B2 level, they are considered to be unsuccessful at the exam (see section 3.2.2 for details regarding TYS).

Validation study 2 explores the potential of using the newly developed Turkish C-test as a screening (readiness) test for the TYS to predict whether candidates are at least at B2 level to be successful in the TYS. A sub-aim is to predict all four TYS levels based on C-test scores since different institutions have different requirements for entry or applications of international students. Therefore, study 2 investigates the predictive power of the Turkish C-test for TYS in educational programs. In contrast to this, validation study 1 did not look at the relation of the C-test with a standardised proficiency test and evaluated the C-test as an instrument that can be used in SLA research studies. By using Turkish C-test as a screening test for TYS, stakeholders can save time, money, and energy before students are committed to take expensive and time-consuming TYS. For this reason, this study involves Turkish instructors as well as TYS candidates to investigate stakeholders' perception towards Turkish C-test.

Similar to validation study 1 in Chapter 6, this chapter starts with the

interpretive argument, and then the validity argument follows in line with Kane's argument-based approach (see section 2.4.4 for details about argument-based approach). Validity argument involves describing participants, instruments, data collection procedures, data analysis methods as well as reporting results and discussing whether the assumptions are supported or rebutted.

7.2 Interpretive Argument of Validation Study 2

The interpretive argument of the validation study 2 is that the Turkish C-test can be used to predict Turkish L2 learners' general language proficiency levels in the TYS before entry to a Turkish-medium university. The inferences, assumptions, and evaluation questions of this interpretive argument are stated in Table 34 below, which was developed by following Kane's (2006) argument-based approach framework.

Table 34. Interpretive argument of validation study 2

Assumptions	Evaluation Questions
Theoretical grounds	
1. The common core of general language proficiency is inclusive of, but not limited to, grammar and lexis.	1. What are the components of general language proficiency?
2. C-tests can assess general language proficiency as demonstrated by a considerable amount of literature.	2. What is the evidence showing that C-test can quickly assess general language proficiency?
3. TYS is a standardized test of Turkish language proficiency aligned with CEFR.	3. What is the structure of the TYS?
Scoring	
4. Text selection and word deletion procedures are appropriate to cover a range of L2 learners in terms of Turkish language abilities.	4. To what extent does the text selection and word deletion procedures produce a test that can cover a range of Turkish L2 learners?
5. Psychometric characteristics of texts are calculated, and all texts have good item fit statistics.	5. Do all texts fit the overall pattern expected by the measurement model?

6. The scoring criteria are appropriate for the test.	6. Are the scoring criteria appropriate? (A2.2)
7. The scoring criteria are applied accurately and consistently.	7. Are the scoring criteria applied accurately and consistently?
Generalization	
8. The C-test texts are internally consistent, and they provide reliable estimates of test takers' L2 abilities.	8. To what extent does the C-test provide reliable estimates of test taker's L2 abilities?
9. The sample of observations is large enough to control sampling error.	9. Is the sample of observations large enough to control for sampling error?
Extrapolation	
10. The C-test scores positively correlate with self-perceived proficiency in Turkish	10. Are there positive correlations between C-test scores and self-perceived proficiency in Turkish?
11. The C-test scores positively correlate with TYS reading, listening, writing, and speaking scores as well as TYS total score.	11. Are there positive correlations between C-test scores and TYS reading, listening, writing, and speaking scores as well as TYS total score?
12. The C-test scores positively correlate with TYS levels.	12. Are there positive correlations between C-test scores and TYS levels?
13. The C-test scores can predict TYS levels and total scores.	13. Can C-test scores predict TYS levels and total scores?
14. The identified C-test cut scores are accurate and sufficient to predict TYS levels.	14. Which C-test cut scores predict TYS levels most accurately and sufficiently?
Decision	
15. The Turkish C-test is useful for TYS candidates to predict their TYS levels and practice their Turkish before taking TYS.	15. What are the perceptions of stakeholders involving instructors of Turkish and TYS candidates regarding the usefulness of the Turkish C-test to predict TYS levels and practice Turkish?

For screening tests, the important features of score interpretations are their meaningfulness and sufficiency for classifying students into levels so that

classification errors will be minimalized (Schmidgall et al, 2017). This is why validation study 2 focuses on extrapolation and decision because it investigates the use of the C-test as a screening test for an external criterion test (TYS) to the contrary of validation study 1 which focuses on scoring and generalization.

7.3 Participants

The validation study 2 involved two types of stakeholders who would benefit from using the Turkish C-test: Turkish L2 learners and instructors of Turkish. This section presents the demographic information of these participants.

7.3.1 Turkish L2 leaners

A total of 79 people out of 3,477 TYS candidates who took the TYS in 2018 or January 2019 participated in the validation study 2 (N=24 January 2018, N=27 July 2018, N=22 May 2018, N=6 January 2019)[26].

Of these learners, 50 were female, 28 were male, and 1 preferred not to state their gender. Since the TYS is a worldwide test, participants were recruited worldwide as well. The highest number of participants were from Turkey (N=26) and Turkey's neighbour country Azerbaijan (N=26) which were followed by Japan (N=4), Kazakhstan (N=4), Bosnia and Herzegovina (N=3), Albania (N=2), Iran (N=2), and Macedonia (N=2). There was also 1 participant from each of the following countries: Austria, Canada, Egypt, Georgia, Kosovo, Lebanon, Poland, Romania, Russia, and Somalia. Overall, participation was mostly from Turkey and countries that are geographically close to Turkey. Given that the highest number of TYS candidates are

[26] Data collection was done between July 2018 and January 2019. Although participants' proficiency levels might have changed between the time they took TYS and Turkish C-test to some extent, it is worth noting that TYS results are valid for two years.

from Azerbaijan, Turkey, and Kazakhstan (see section 3.2.2), the participants are predicted to represent where most TYS candidates are from.

In order to form a heterogenous sample in terms of the language proficiency, all levels involving unsuccessful TYS candidates (below B2 level) were invited in the study. However, the number of participants across different TYS levels was not equal, and there were only 5 participants who were unsuccessful at TYS as can be seen in Table 35. Since no information could be reached about the actual level distribution of all people who have taken TYS up until today although it was requested, it is not known how representative this sample is in terms of the TYS levels.

Table 35. Distribution of participants according to TYS levels

TYS level	Below B2	B2	C1	C2	Total
N	5	13	50	11	79

The age of the participants ranged between 16 and 46 with a mean of 23.84. There was one usual participant at the age of 46. However, most of the participants were between 16 and 35 years old, showing the typical age for university studies.

In relation to the completed degree of education, the majority of the participants had a bachelor's degree (N=41), followed by a high school degree and a master's degree at equal numbers (N=18), and a doctoral degree (N=2). The subject of the completed degree of education varied from social sciences to positive sciences. However, the most common subjects were language and literature studies (N=11) followed by biology and chemistry (N=7), engineering (N=7) and medicine (N=6).

Regarding L1, the sample was heterogenous with 18 different L1s in total. Nevertheless, 45 participants had a Turkic L1 (i.e., Azerbaijani, Kazakh) which comes

from the same language family as Turkish. There were 39 Azerbaijan[27]i, 5 Bosnian, 5 Japanese, 4 Albanian, 4 Arabic, 4 Kazakh, 3 Persian, 3 Georgian, 2 English, and 2 Russian L1 speakers. There was also 1 L1 speaker of each of the following languages: German, Polish, Romanian, Somali, Spanish, Tatar, Turkish[28], and Uyghur. Furthermore, 39 participants self-identified as heritage speakers of Turkish, and 26 participants identified themselves bilingual speakers of Turkish on the survey. These groups were not investigated as separate groups since TYS scores were taken as the main criterion to evaluate Turkish C-test scores.

The Turkish language background information of the participants is shown in Table 36 below. As can be seen, the sample generally seems to have lots of exposure of Turkish language, which aligns with the fact that 94% of participants have been successful at TYS as shown in Table 35 above.

Table 36. Participants' Turkish background information

Variable	Mean	SD	Min	Max
Age of learning	14.90	8.09	0	34
Months of study	81.54	71.05	2	252
Months of residence in Turkey	14.55	35.94	0	192
Hours of study per week	14.40	24.92	0	160

[27] Although the study was conducted worldwide, most takers had Azerbaijani as their L1 reflecting the population characteristics for TYS candidates (see section 3.2.2). Azerbaijani belongs to the same language family as Turkish and there are many similarities between these two languages. Thus, the factor of L1 background through DIF analysis was not investigated since it is expected that Azerbaijani learners could possibly do better on the test, which does not mean there is a bias in the test against non-Azerbaijani speakers.

[28] This participant was included since she self-identified as a Turkish heritage speaker and her TYS level was B2.

Thirteen of the learners volunteered to participate in a follow-up interview. Interviews lasted between 12 and 29 minutes. Seven of the interviews were conducted in Turkish upon interviewees' request. The details of this group of learner interviewees are given in Table 37 below. Note that there is no interviewee below B2 level.

Table 37. Learner Interviewee Data

ID	Gender	Occupation (age in brackets)	L1	TYS level	TYS total score	C-test score (over 160)	Reason for taking TYS
1	M	student (24)	English	C1	82	120	future employment and see his own progress
7	M	economist (33)	Kazakh	C1	83.5	135	future employment and see his level of Turkish
8	F	student (24)	Azeri	C2	93.05	125	postgraduate study in Turkey
9	F	student (24)	Azeri	B2	69.79	131	postgraduate study in Turkey
10	F	data analyst (32)	Romanian	C1	82.28	100	future employment
16	M	student (19)	Azeri	C1	75	95	study in Turkey
21	F	student (17)	Azeri	C1	84.91	99	study in Turkey
25	F	researcher (28)	English	C1	74.96	103	application for PhD with a focus on Turkey
30	F	receptionist (24)	Russian	C1	76.41	120	currently in Turkey and wants to make applications to Turkish universities
31	M	graphic designer(46)	Spanish	C1	85	87	his current employment in embassy
35	M	student (22)	Azeri	C1	84.13	114	postgraduate study in Turkey

38	M	student (22)	Japanese	B2	60.28	85	just from his curiosity[29]
69	M	computer engineer (37)	Russian	C1	81	137	currently in Turkey and looking for a job change

7.3.2 Instructors of Turkish

A total of 34 instructors of Turkish from 16 different countries participated in the validation study 2 and completed an instructor survey. These instructors were mostly from Turkey (N=12), which was followed by Bosnia and Herzegovina (N=4), Albania (N=2), Azerbaijan (N=2), Georgia (N=2), and Iran (N=2). Also, there was 1 instructor from each of the following countries: Afghanistan, Austria, Germany, Jordan, Kazakhstan, Kosovo, Malesia, Serbia, Tunisia, and USA. All these instructors were Turkish L1 speakers.

The number of female and male instructors were equal. Their age ranged between 23 and 40 with a mean of 29.74. All of them, except one, heard or read about TYS before. Along a similar line, 21 of them heard or read about C-test before, and 11 of them used a C-test either in English or Turkish as a teaching tool.

Two of the instructors volunteered to participate in a follow-up interview (see section 7.8 for a discussion of having few instructor interviewees). The first interviewee was a female language instructor aged 26 from Turkey, and the second interviewee was a male language instructor aged 30 from Turkey. Both interviews lasted between 15 and 20 minutes and were conducted in English.

[29] This participant studied in an English-medium university in Turkey for a semester with an exchange program. So, he wanted to see how much he could manage in Turkish without problems.

7.4 Instruments

Instruments for Turkish L2 learners involve background questionnaire, Turkish C-test, feedback survey, and interview questions. Also, instruments for instructors involve interview questions and surveys.

7.4.1 Background Questionnaire

An online background questionnaire was administered to learners before they took the Turkish C-test to find out: (1) participant demographic information for generalization inference; (2) TYS results for extrapolation inference. Although the background questionnaire as well as the survey and the test instructions were initially only in English, they were also translated to Turkish after the first ten participants upon seeing that some participants with L1 other than English (i.e., Azerbaijani L1 speakers) had problems with answering the open-ended questions. Overall, 60 participants completed the study in Turkish while 19 did it in English.

The questionnaire involved questions about general demographic information, Turkish learning history and use, and self-perceived proficiency in Turkish. It was the slightly revised version of the background questionnaire used in validation study 1 (see section 6.4.1). The revision included two extra questions: (1) TYS level, total score, and scores in reading, listening, speaking, and writing sections; (2) the location and the time that TYS was taken (see Appendix 18) .

7.4.2 The C-test

After the background questionnaire, the best functioning 8 texts from the development stage of the Turkish C-test were administered to participants online (see Appendix 19 for the 8-text Turkish C-test and Chapter 5 for the test development stage). The reason why there were 8 texts in study 2 is that an attempt was made to have 2 texts per TYS level given that there are four levels at TYS (below B2, B2, C1, C2). This is similar to

the onSET, an 8-text C-test which has two texts per level and is used as a screening test for the TestDaF (i.e., Eckes, 2014). In contrast, in study 1, the test aimed to be used as an SLA research tool to distribute learners along a range of scores, and thus one text per level was kept in each grouping of 5 different levels of difficulty aligning with studies using C-tests for research purposes (i.e., Lee-Ellis, 2009; Norris, 2018).

Participants were given a total of 40 minutes (5 minutes per text) to complete the test once they started it since the test was unsupervised. They could choose Turkish special characters (ç, ı, ğ, ö, ü, ş) from a text box when they were completing the gaps. They were warned not to use any external aids and to be careful about spelling since spelling counted. Texts were ordered according to their difficulty with Text 1 being the easiest and Text 12 being the most difficult in order to facilitate the familiarization of learners with the test format. Table 38 below shows the details about each of the texts. Note that the same texts are given the same text numbers across the three empirical chapters to make the comparison easier later in the discussion.

Table 38. Level and characteristics of the 8-text C-test

Text	**ILR Level**	**Topic**	**Characteristics**	**Source**
1	1	Locations	Very basic sentences with "there is/there are" structure. Familiar words, cognates.	Created based on commercial textbooks
3	1	A Danish person in Turkey	Simple sentences with present continuous. Concrete words, some cohesive device.	Graded Turkish reader book (authentic texts)
4	1+	Description of a Turkish city	Simple sentences with relative clauses. Informative social purpose.	Adapted from an airline website
6	1+	The advertisement of a school	Simple and compound sentences. Concrete words, a few abstract words.	A School website

7	2	Student success	Cause and effect relations. Factual information.	A health organization website
9	2+	Relation between taste and smell	Conditionals and negations. Topic specific vocabulary.	Newspaper
11	3	Turkish science women	Social issue, abstract topic. General report. Evaluative statements.	Newspaper editorial column
12	3+	The relation between cultural venues and folk dances.	Social and abstract topic. Less-frequently used and more advanced words. Long and complex sentence structures.	Academic journal in social sciences

7.4.3 Test Taker Feedback Survey

After completing the test, participants were asked to complete the test taker feedback survey to get learner input for the decision inference. The survey involved questions about participants' test taking experience and views about the Turkish C-test such as *Was there anything that confused you while completing the test and the questionnaire?* and *If you found some texts more difficult than others write the reasons why you found them more difficult*. It was the revised version of the feedback survey used in study 1 (see section 6.4.3). The revisions included the following: (1) a Likert-scale question about to what extent test takers think the Turkish C-test is useful to practice language skills before taking TYS and an open-ended question about why they think so; (2) a Likert-scale question about to what extent test takers think the Turkish C-test is useful to quickly estimate TYS level and an open-ended question about why they think so (see Appendix 20). At the end of the survey, participants were asked whether they would like to be reimbursed for their participation with Starbucks e-gift cards or Idefix (bookstore) e-gift cards and whether they would like

to participate in a follow-up online interview. If they preferred to participate in the interview, they were asked whether they would prefer a video or a phone call and select an interview date on the embedded calendar which was created by using Calendly[30].

7.4.4 Interview Questions for Test Takers

Semi-structured interviews were conducted with 13 test takers who expressed an interest in a follow-up interview. All interviews were conducted via Skype using Ecamm Call Recorder for Skype. The aim of the interview was to ask test takers to elaborate on their responses to the survey questions about their views on the C-test and TYS. The interviews lasted between 12 and 30 minutes. The questions were related to overall Turkish language learning experience (i.e., whether they can reach enough practice materials), TYS experience (i.e., the reason why they took TYS and their impression of TYS), and Turkish C-test experience (i.e., the difficulty level, the comparison of Turkish C-test and TYS). For example, there were questions such as *How would you compare your performances in TYS and Turkish C-test* and *What was your impression of Turkish C-test? How well did you do in Turkish C-test?* (see Appendix 23 for the interview questions for test takers).

7.4.5 Survey for Instructors

An online survey was administered to instructors of Turkish in order to investigate whether the Turkish C-test is a useful tool for instructors by eliciting their perception of the test for the decision inference. The survey included questions about demographic information, and instructors' familiarity with and views about C-test. It was a revised version of the survey for researchers used in study 1 (see section 6.4.4).

30 https://calendly.com/

The revisions included the following: (1) yes/no and short answer questions about researchers' familiarity and experience with TYS, as well as a Likert-scale question related to instructors' views about the usefulness and fairness of TYS; (2) Likert-scale and open-ended questions related to researchers' views about the usefulness of Turkish C-test to estimate TYS levels (see Appendix 21 for the survey for instructors). Instructors read an overview of C-tests and TYS as well as the instructions and texts of the Turkish C-test used in the study. At the end of the survey, they were asked whether they would like to be reimbursed for their participation with Starbucks or Idefix (bookstore) e-gift cards and whether they would like to participate in a follow-up online interview.

7.4.6 Interview Questions for Instructors

Semi-structured interviews were conducted with 2 instructors of Turkish who wished to participate in a follow-up interview. One of the interviews was conducted via Skype using Ecamm Call Recorder for Skype. The other interview was conducted via Zoom because the interviewee preferred that option. The aim of the interview was to ask instructors to elaborate on their responses to the survey questions about their views on the C-test and TYS. The interviews lasted between 15 and 20 minutes. The questions were related to demographic information (i.e., work experience, professional background), views of the TYS and Turkish C-test (i.e., appropriateness for students) as well as the interpretation of the relationship between TYS and Turkish C-test. For example, there were questions such as *Do you think Turkish C-test can be used as predictive of student performance on Turkish Proficiency Exam? Why or why not?* (see Appendix 22 for interview questions for instructors).

7.5 Data Collection Procedures

TYS test takers and instructors were recruited through e-mail invitations which were sent to the headquarters of YEE in Ankara. After the official permission for study distribution was taken from the director of the YEE, the head office of YEE distributed the e-mail invitations to their branches worldwide to be shared with test takers who took TYS between January 2018 and January 2019. In addition to this, e-mail invitations were also sent to instructors of Turkish and mailing lists (i.e., American Association of Teachers of Turkic languages). A participation invitation with an overview of the research was also published in the January 2018 newsletter of the American Association of Teachers of Turkic Languages.

The online platforms used for questionnaire, survey, and C-test were the same as in study 1. Qualtrics was used for background questionnaire and survey while Learnclick was preferred for C-test due to the test format (see section 6.5 for details and see Appendix 24, 25, 26, 27 for test taker, instructor, interviewee test taker, and interviewee instructor information sheet and consent forms in turn). Participants were able to participate in the study by anytime anywhere they wished as long as they had internet connection.

The average time participants spent on the 8-text C-test was found to be 30 minutes. Test takers were shown their total score and score percentage at the end of the test as in study 1.

7.6 Data Analysis Methods

Following are the methods which were used to obtain required evidence to support each inference by answering the related EQs. As mentioned in the previous validation study 1 in Chapter 6, the analysis of the theoretical grounds inference was not

included since it is based on literature review and the results relating to theoretical grounds are reported in results of theoretical grounds section 7.7.1.

7.6.1 Analysis of Scoring Inference

This section relates to EQs 4-7 under the scoring inference as presented in Table 39. These evaluation questions were also used in the validation study 1 except a slight difference in EQ 5 where the aim is to investigate whether all texts have good item statistics rather than choosing the best functioning five texts for practicality. Therefore, the analyses of the questions were done in the same way (see section 6.6.1).

Table 39. Scoring Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods
4. Text selection and word deletion procedures are appropriate to cover a range of L2 learners in terms of Turkish language abilities.	4. To what extent does the text selection and word deletion procedures produce a test that can cover a range of Turkish L2 learners?	4. Expert judgement, Rasch analysis (examinee separation indices), teacher perception
5. Psychometric characteristics of texts are calculated, and all texts have good item fit statistics.	5. Do all texts fit the overall pattern expected by the measurement model?	5. Rasch analysis (item fit indices, item discrimination values, item difficulty measures, standard error estimates), Item-examinee maps
6. The scoring criteria are appropriate for the test.	6. Are the scoring criteria appropriate?	6. Answer key based on undeleted versions of the words and alternative answers, TUD
7. The scoring criteria are applied accurately and consistently	7. Are the scoring criteria applied accurately and consistently?	7. Automatic scoring

7.6.2 Analysis of Generalization Inference

This section relates to EQs 8 and 9 under the generalization inference as presented on Table 40. These evaluation questions were also included in the validation study 1; therefore, they were analysed in the same way (see section 6.6.2).

Table 40. Generalization Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods
8. The C-test texts are internally consistent, and they provide reliable estimates of test takers' L2 abilities.	8. To what extent does the C-test provide reliable estimates of test taker's L2 abilities?	8. Reliability analysis
9. The sample of observations is large enough to control sampling error.	9. Is the sample of observations large enough to control for sampling error?	9. Item-person maps, separation indices, literature review

7.6.3 Analysis of Extrapolation Inference

This section relates to EQs 10 to 14 under the extrapolation inference as presented in Table 41. As mentioned in section 7.2., validation study 2 focuses on extrapolation since it investigates the use of the Turkish C-test as a screening test for the TYS, and this is where the differences really lie between study 1 and study 2.

Table 41. Extrapolation Inference Assumptions and Evaluation Questions

Assumptions	Evaluation Questions	Methods
10. The C-test scores positively correlate with self-perceived proficiency in Turkish.	10. Are there positive correlations between C-test scores and self-perceived proficiency in Turkish?	10. Spearman's rho correlation coefficient
11. The C-test scores positively correlate with TYS reading, listening, writing, and speaking scores as well as TYS total score.	11. Are there positive correlations between C-test scores and TYS reading, listening, writing, and speaking scores as well as TYS total score?	11. Spearman's rho correlation coefficient
		12. Spearman's rho correlation coefficient

12. The C-test scores positively correlate with TYS levels.	12. Are there positive correlations between C-test scores and TYS levels?	13. Ordinal logistic regression, linear regression
13. The C-test scores can predict TYS levels and total scores.	13. Can C-test scores predict TYS levels and total scores?	14. Classification table
14. The identified C-test cut scores are accurate and sufficient to predict TYS levels.	14. Which C-test cut scores predict TYS levels most accurately and sufficiently?	

Correlational analyses were conducted between C-test scores and several indicators of proficiency including self-perceived proficiency, TYS scores and level. Regarding EQ10, Spearman's rho was calculated between C-test scores and self-perceived proficiency estimates on a 5-point Likert-scale due to the ordinal nature of the Likert scale. In a similar way, regarding EQs 11 and 12, Spearman's rho was calculated between C-test scores, TYS levels as well as TYS scores in reading, writing, listening, speaking and TYS overall score due to the ordinal nature of TYS levels and non-normality of C-test and TYS scores. However, since correlations between C-Test and TYS would not be enough to justify the predictive power of C-test for TYS levels and scores, first ordinal logistics regression and then linear regression were conducted to predict possibilities of attaining TYS scores and each categorical TYS level as a smooth function of polynomial C-test scores addressing EQ 13. Regarding EQ 14, a classification table was created by cross tabulating the levels predicted by C-test (derived from regression) and observed levels, which allowed to reveal how many points a test taker should achieve on C-test to get a certain level on TYS.

7.6.4 Analysis of the Decision Inference

This section relates to EQ 15 under the decision inference as presented on Table 42. A similar EQ was included in study 1 regarding the perception of stakeholders under the decision inference. Thus, the analysis was conducted in the same way as in study 1 (see section 6.6.4)

Table 42. Decision Inference Assumption and Evaluation Question

Assumptions	Evaluation Questions	Method
15. The Turkish C-test is useful for TYS candidates to predict their TYS levels and practice their Turkish before taking TYS.	15. What are the perceptions of stakeholders involving instructors of Turkish and TYS candidates regarding the usefulness of the Turkish C-test to predict TYS levels and practice Turkish?	15. Thematic analysis of interviews and open-ended survey questions, descriptive statistics of Likert-scale survey questions

7.7 Results

This section reports the results of the data analysis explained in the previous section 7.6. It presents results under each specific inference that they relate to (see the next section 7.8 for a discussion of these results).

7.7.1 Results for the Theoretical Grounds Inference

This part presents the evidence for the theoretical grounds, which is based on the literature review. EQ 1 *"What are the components of general language proficiency"* and EQ 2 *"What is the evidence showing that C-test can quickly assess general language proficiency?"* are the same as in the validation study 1; therefore, see section 6.7.1 for the results. For EQ3 *"What is the structure of the TYS?"*, the TYS is a standardized test of language proficiency involving four main language skills and aligning with CEFR (see section 3.2.2 for details about TYS). It is used to assess the

language proficiency of Turkish L2 learners for their registration at Turkish-medium universities or employment at the governmental level.

7.7.2 Results for the Scoring Inference

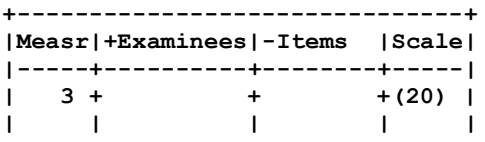
7.7.2.1 Text Selection and Word Deletion (EQ4)

EQ 4 relates back to the test development stage in Chapter 5 (see section 5.5.6.1 for details) as well as analyses conducted in this chapter. Expert judgement showed that the 8-text C-test ranged between ILR 1 and ILR 3+ levels. In addition to this, teachers' perception of the text difficulty corresponded with the ILR ratings of each text as seen on Table 43 below. The mean of perceived text difficulty level (1- the easiest, 5- the most difficult) gets gradually higher from text 1 towards text 12, except for Text 7 which was found to have repetitive words.

Table 43. Instructors' perception of text difficulty (N=34)

	Text 1 ILR 1	Text 3 ILR 1	Text 4 ILR 1+	Text 6 ILR 1+	Text 7 ILR 2	Text 9 ILR 2+	Text 11 ILR 3	Text 12 ILR 3+
Mean	2.00	2.41	2.91	3.24	3.15	3.32	3.59	4.03
Min	1	1	1	1	1	1	1	1
Max	4	4	4	5	5	5	5	5

The variance in total scores of 8 text C-test explained by the Rasch model was 81.63%. It was able distinguish at least 3 different levels of examinees with .92 reliability (Examinee Separation= 3.51; Strata= 5.01). Figure 26 below shows the item-examinee map. Higher level examinees and more difficult texts are displayed at the top of the figure while lower level examinees and less difficult texts are displayed at the bottom.



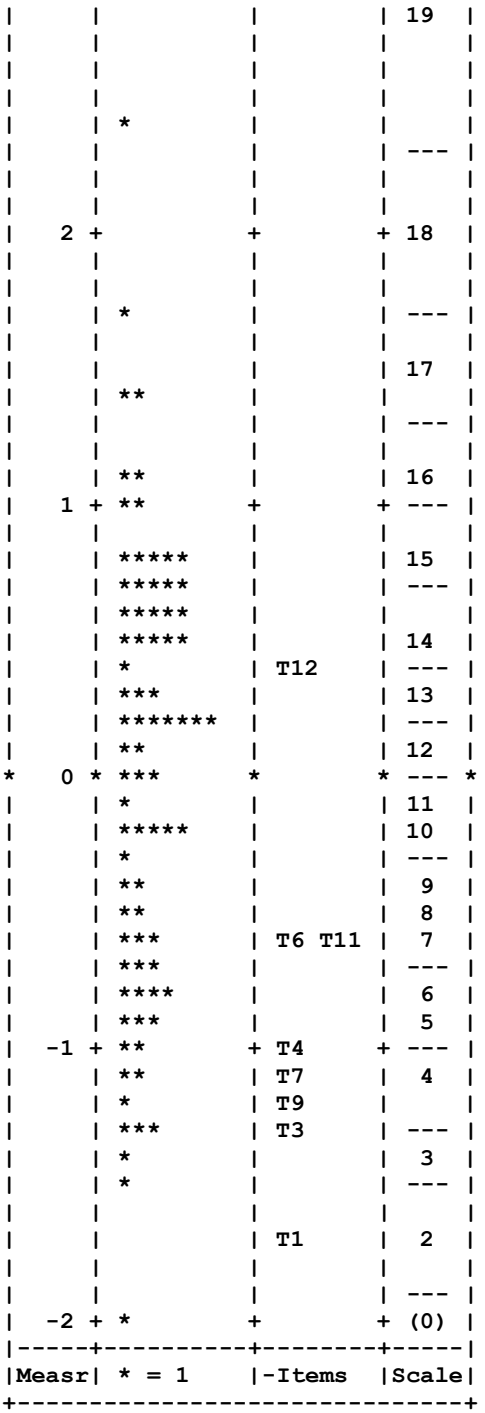


Figure 26 . Item-person Map for 8-text C-test (N=79)

Looking at Figure 26, it was revealed that T1 didn't contribute much to the overall test by being too easy for the sample. T9 also surprisingly turned out to be easy for the sample of study 2 although it was one of the most difficult texts for the samples in the test development stage and study 1. One explanation might be that Text 9 might have many cognate words with Azerbaijani and the 49.4% of the sample

for study 2 consisted of Azerbaijani L1 speakers while 62.1% of the sample for study 1 consisted of English L1 speakers. Another possibility might be that T9 was about the relation between smell and taste, and 16% of the sample was working in the fields of medicine (N=6) or biology and chemistry (N=7). For example, in the interview, one examinee commented that she is a chemist teacher, so she might be more familiar with that topic and related vocabulary in Turkish. Nevertheless, Text 1 and Text 9 were not removed from the study since they had acceptable item fit statistics addressing the assumption.

As seen in Figure 26, although most of the test takers were covered between Text 1 and Text 12, there was a small group of test takers above the difficulty level of the most difficult Text 12. Nevertheless, given the screening purpose of the test, there is no need for texts distinguishing among high level learners. Rather, texts around the decision point are more important. However, if future test users may want to use the test for distinguishing among high-level learners, they should replace the easiest text(s) (i.e., Text 1 which did not seem to contribute to the overall difficulty) with a more difficult one (see section 7.8 for discussion).

7.7.2.2 Psychometric Characteristics of C-test Texts (EQ5)

In order to answer the EQ 5 about the texts fitting the expected model, Rasch analysis was conducted by using 2-facet (examinee + item) RSM (see section 4.5.2.1 for an explanation of RSM). Table 44 shows the psychometric characteristics of the 8 texts from the present chapter after the first analysis (see section 4.5.2.2.2 for an explanation of these item statistics).

Table 44. Key item quality statistics of 8 texts (N=79)

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.74	1.04	.91	.96	.08	-1.74
T3	.75	1.11	.89	.85	.07	-1.33

T4	.73	.82	1.36	1.14	.07	-.98
T6	.82	1.14	.91	.91	.05	-.64
T7	.83	1.29	.73	.71	.06	-1.10
T9	.82	.83	1.36	1.21	.07	-1.18
T11	.81	.95	1.06	1.01	.06	-.60
T12	.83	1.17	.73	.72	.05	.42

All texts submitted acceptable item fit indices (infit and outfit between .5 and 1.5), that is they performed as expected by the model. The easier texts (T1, T3, T4) had the lowest point biserial values (Rpbi) under .80, which meant the scores from these texts correlated the least with the total score. Regarding discrimination values, T4 and T9 had the lowest discrimination values, but they were still within the acceptable range of .5 and .15. The difficulty measures aligned with the item-examinee map on Figure 26 above, and there was only one text with a positive difficulty measure.

In addition to item quality statistics, four outlier examinees (E16, E39, E43, E69) were identified in terms of their performance since their infit/outfit statistics were greater than 2.0 after the first analysis (see section 4.5.2.2.1 for an explanation of examinee statistics). Looking closely at the texts these outliers completed, it was found that two of these four outliers had left some texts empty despite performing well on the other texts: E39 had left T4 empty, and E43 had left the last two texts empty. On the other hand, E16 left many of the blanks empty on the easier texts, and E69 got lower points from the easier texts at the beginning of the test while having 19 point from the most difficult last text. Evidencing this result, in the semi-structured interview, E69 revealed that he did not pay much attention while completing the first few texts as can be seen in the excerpt below:

> I wasn't very attentive in the first several texts. I didn't even think. I could easily guess the word. So, I wrote the first thing that came to my mind, and probably I wasn't very attentive to the grammar. But, if I reread, maybe I would score better. For the last two texts, I had to reread several times.

So, it seems that these test takers did not try equally hard on all texts. As also explained in study 1 (see section 6.7.2.2), results would slightly improve if these outliers were removed from the analysis (see Appendix 28). However, given the small sample size and considering that individuals like this are part of the population to which we wish to generalize, it can also be interpreted as adapting the sample to match the model. Therefore, these outliers were kept.

7.7.2.3 Appropriateness and Accuracy of the Scoring Criteria (EQ6 and EQ7)

Regarding the EQ 6 related to the appropriateness of the scoring criteria, the answer key including undeleted versions of the words and alternative answers proved to be appropriate to score the test taker answers. In relation to the EQ 7 about the application of scoring criteria, automatic scoring by Learnclick enabled accuracy, consistency, and objective results.

7.7.3 Results for the Generalization Inference

7.7.3.1 Reliability of the C-test (EQ 8)

A reliability test was conducted for the 8-text C-test addressing the EQ8, and a Cronbach's alpha of .95 was found for the 8-text C-test.

7.7.3.4 Sufficiency of the sample size (EQ9)

Sufficiency of the sample size was justified in the same way as in study 1 (see section 6.7.3.4). To briefly summarise, the sample size of this validation study 2 (N=79) is aligned with other validation studies in LCTL which used IRT analysis as well as inferential statistics (i.e., Drackert, 2016; Son, 2018). Furthermore, IRT examinee separation index of 3.51 and item-examinee map showed that the test was able to distinguish at least 3 different levels of learners involved in the sample.

7.7.4 Results for the Extrapolation Inference

This section reports the relation between C-test scores and several other indicators of language proficiency involving self-perceived proficiency, TYS levels and TYS scores.

7.7.4.1 Correlations between C-test scores and self-perceived proficiency (EQ 10)

Before addressing the EQ 10 about the correlation between C-test scores and self-perceived proficiency, the descriptive statistics of the ordinal self-perceived proficiency scores are provided in Table 45 below.

Table 45. Descriptive statistics of self-perceived proficiency (N=79)

	self-reading	self- listening	self-writing	self- speaking	self-overall
Mean	4.27	3.75	4.14	3.92	3.96
Median	4	4	4	4	4
Mode	5	4	4	4	4
Min	1	1	1	1	1
Max	5	5	5	5	5

Following this, Spearman's rank-order correlation coefficient (Spearman's rho) was calculated due to the ordinal structure of scores in Likert-scale self-perceived proficiency. The highest correlation of C-test scores was with the self-perceived proficiency in reading skill, followed by overall, speaking, listening, and writing skills as seen in Table 46.

Table 46. Spearman's rho between C-test and self-perceived proficiency (N=79)

	Turkish C-test (8-text)
Self-reading	.50
Self-listening	.41
Self-writing	.26
Self-speaking	.43
Self-overall	.47

Note: all correlations statistically significant, $p < .005$

All correlations were over .40, except the lower correlation with writing (see section 7.8 for discussion). The correlations between self-perceived proficiency and

each TYS skill were also calculated and provided in Appendix 29 for further insight into low correlation with writing although it is not directly answering the EQ. Self-perceived writing proficiency had lower correlation with TYS skills sections.

7.7.4.2 Correlations between C-test scores and TYS scores and level (EQ 11&12)

Before conducting correlations and regressions between the two tests, descriptive statistics were calculated for both C-test and TYS scores. Descriptive statistics of the 8-text C-test are shown on Table 47 below.

Table 47. Descriptive statistics of C-test total scores

		8-text C-test
N		79
K		160
Mean		114.53
Std. Error of Mean		2.72
95% confidence interval for mean	Lower bound	109.11
	Upper bound	119.95
Median		122
SD		24.20
Min		40
Max		154
Range		114
Variance		585.38
Skewness		-.788
Std. Error of Skewness		.271
Kurtosis		.010
Std. Error of Kurtosis		.535

As can be seen on Table 47, the C-test elicited scores ranging from 40 to 154 out of a total possible score of 160. The moderate negative skewness and high mean values indicated that more participants were grouped around the higher end of the distribution (see Figure 27 below). Therefore, in order to assess the normality of C-test scores to check whether a Pearson correlation coefficient would potentially be suitable for testing the correlation with the TYS score, the Kolmogorov-Smirnov test for normality was used, and normality was rejected (D (74) = .144, p= .001).

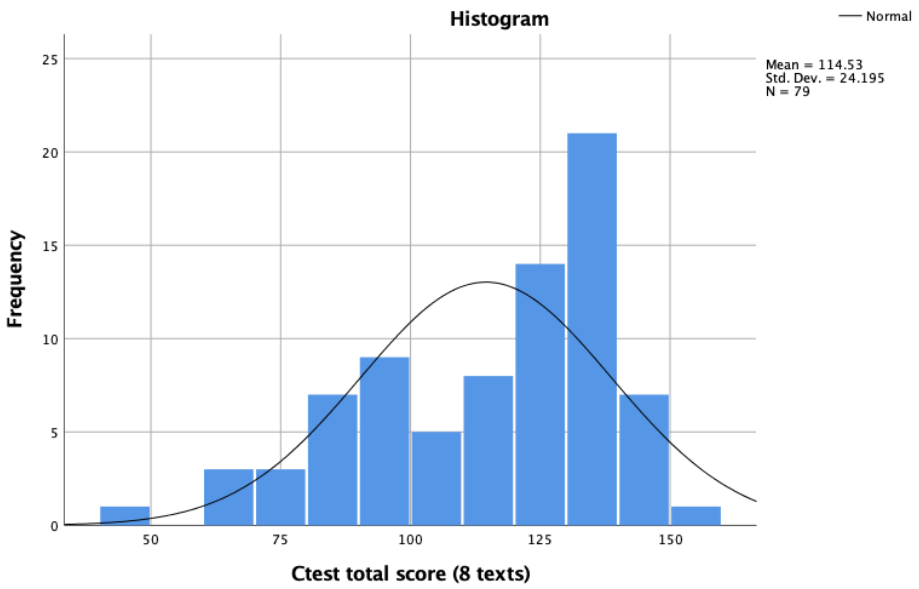


Figure 27. 8-text C-test score distribution (N=79)

Regarding TYS, descriptive statistics of total score and reading, writing, listening, and speaking scores are detailed in Table 48.

Table 48. Descriptive statistics of TYS

	TYS total	TYS reading	TYS listening	TYS writing	TYS speaking
N	79	78	78	78	78
k	100	25	25	25	25
Mean	76.88	20.46	18.59	17.75	20.05
Std. Error of Mean	1.48	.43	.46	.44	.49
95% CI Lower bound	73.94	19.60	17.67	16.87	19.08
Upper bound	79.82	21.31	19.52	18.63	21.02
Median	80.09	21.88	19.17	18.44	20.88
SD	13.04	3.79	4.11	3.91	4.31
Min	17.24	9.38	2.5	3.36	2
Max	94.83	25	25	24.13	25
Range	77.59	15.62	22.5	20.77	23
Variance	170.06	14.4	16.86	15.29	18.55
Skewness	-1.726	-1.153	-.851	-.940	-1.555
Std. Error of Skewness	.272	.272	272	.272	.272
Kurtosis	4.878	.727	1.645	1.266	3.369
Std. Error of Kurtosis	.538	.538	.538	.538	.538

TYS reading skill had the highest mean, which was followed by speaking, listening, and writing. No test taker was also able to get the maximum possible score of 25 from the writing section while they could in other skill sections. Therefore, it seems that test takers had the most difficulty with the writing section. TYS total score

ranged between 17.24 and 94.83 with a mean of 76.88. Remember that TYS levels are based on the total score (B2:55-70; C1: 71-88; C2: 89-100). However, candidates should get a minimum of 12.5 in each skill section in order to get a certificate.

Figure 28 shows that TYS total scores were clustered around the high end of the distribution as indicated by the high negative skewness and high kurtosis. This would not be surprising given that 94% of the participants were successful at TYS as explained in section 7.3.1. The Kolmogorov-Smirnov test of normality confirmed the non-normal distribution of TYS total scores (*D* (78) =131, p=.002).

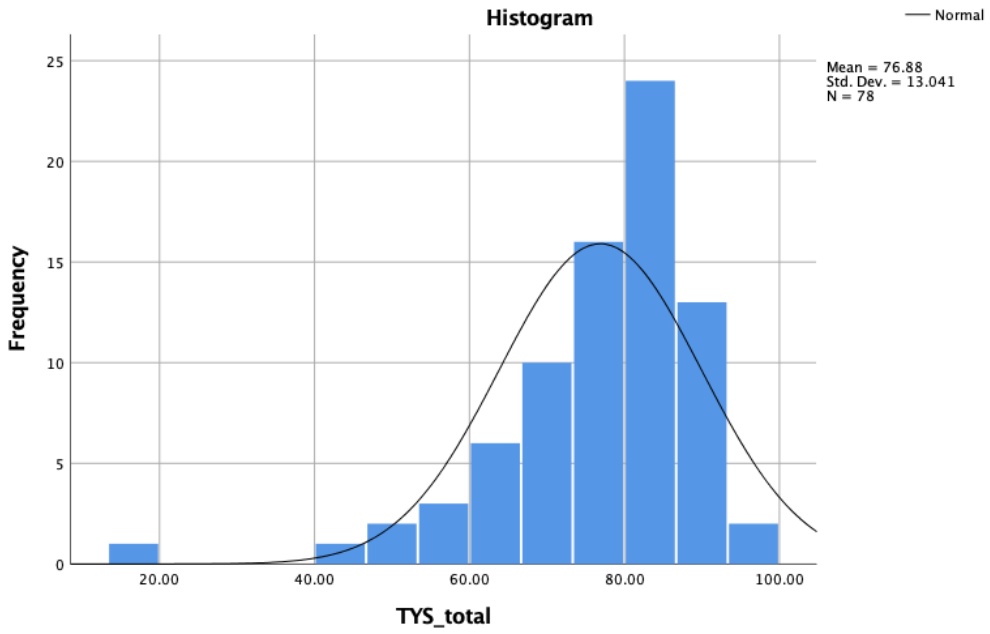


Figure 28. TYS total score distribution (N=79)

Regarding the distribution of scores in TYS skill sections, scores were not normally distributed for TYS reading (*D* (78)=159, *p*<.001) and TYS speaking (*D* (78)=.141, *p*=.001) according the Kolmogorov-Smirnov test. On the other hand, they were normally distributed in TYS listening (*D* (78)=.081, *p*=.051) and TYS writing (*D* (78)= .100, *p*= .200) (see Appendix 30 for distributions of scores in TYS skill sections).

After calculating descriptive statistics and distributions of scores, correlational analyses were conducted. The scatterplot shows a positive relationship between TYS total scores and C-test scores on Figure 29. Examinees below B2 (coloured red), who failed at TYS, are also among the ones who got the lowest scores on C-test. There are more examinees between 120 and 140 points on C-test and between 70 and 90 points on TYS, showing that most of the examinees are clustered around the higher scores.

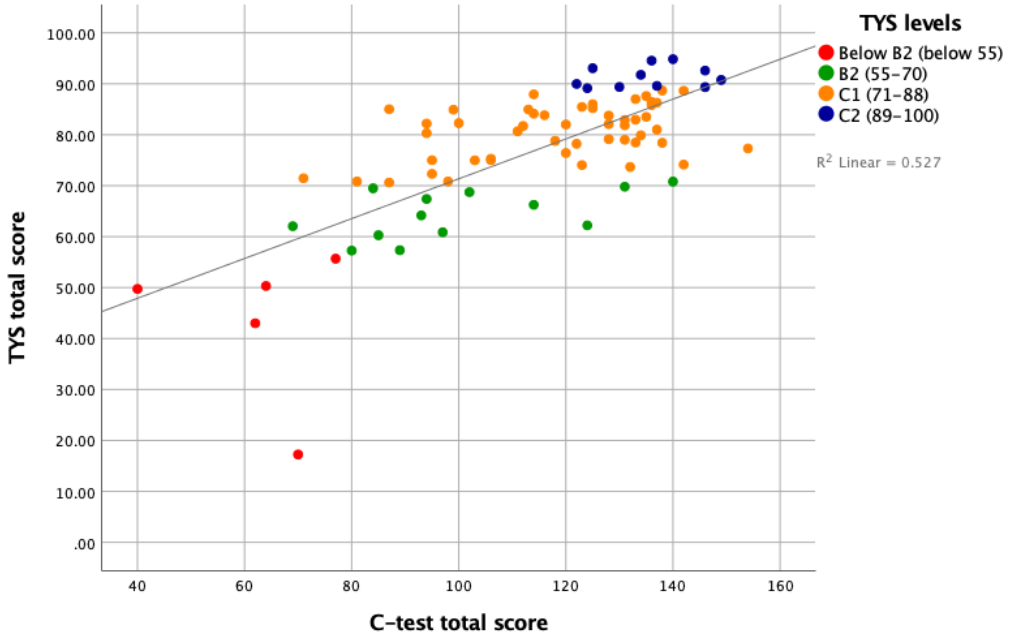


Figure 29. Scatterplot between C-test and TYS total score with linear fit line

Figure 29 shows that the linear regression, where the scores of the C-test and TYS are expected to increase proportionally, does not seem to work so well after a certain level. This is because after a certain score in C-test, TYS total score does not increase so rapidly even though there is an increase in the C-test score. Therefore, a quadratic or cubic line might work better in this relationship as can be seen in Figures 30 and 31.

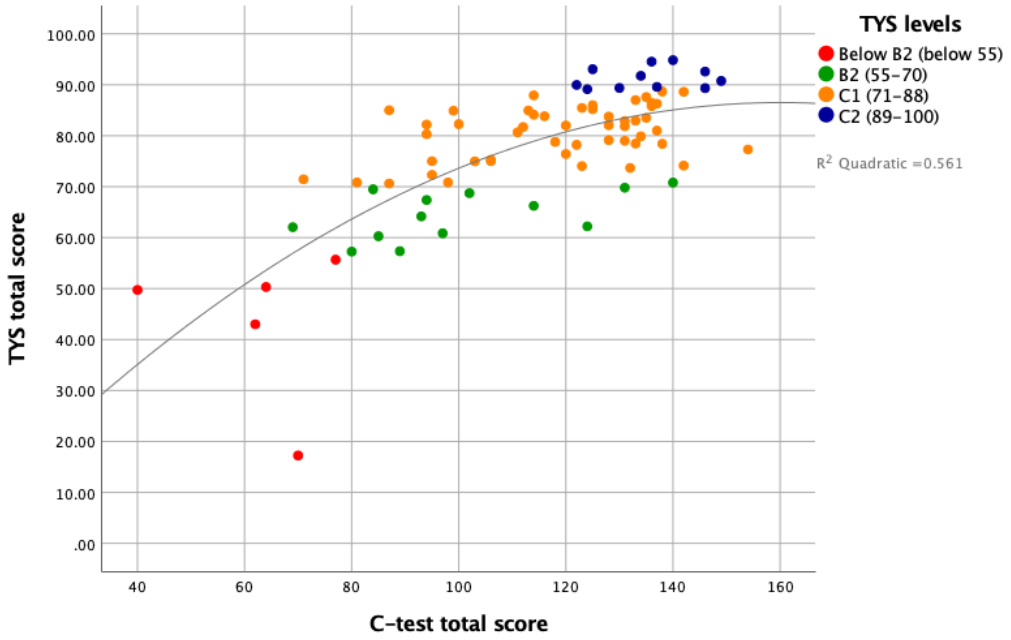


Figure 30. Scatterplot between C-test and TYS total score with quadratic fit line

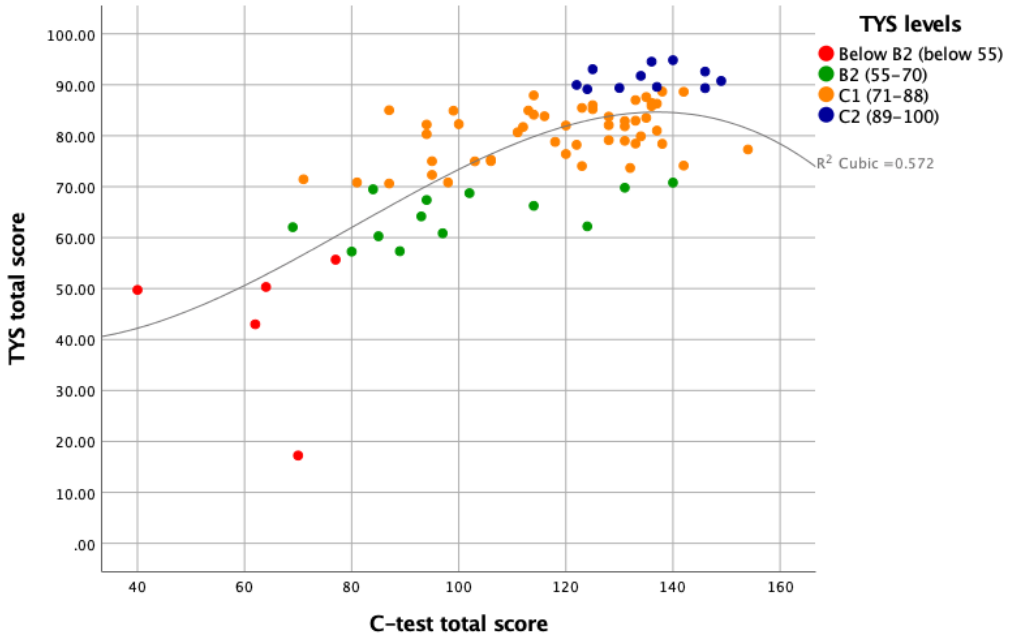


Figure 31. Scatterplot between C-test and TYS total score with cubic fit line

The quadratic and especially the cubic polynomials work less well at the extremes of the C-test score distribution (a known problem with polynomials), but they work fine for the majority of test takers and in the middle of the distribution where the crucial pass-fail threshold of the TYS lies. A quadratic line seems to work better than a cubic line since the cubic line shows a nonsensical reduction in the

predicted TYS score after a C-test score of 145. In contrast, the quadratic line shows a concave relationship. Therefore, the quadratic model is chosen over the cubic model.

Following the examination of scatterplots, Spearman's rho was calculated. As seen on Table 49 below, the C-test correlated with the total score and receptive skills (listening and reading) higher than it correlated with productive skills (writing and speaking). It correlated the most with listening (ρ=.67), followed by total score (ρ=.65), TYS level (ρ=.59), reading (ρ=.58), speaking (ρ=.49), and finally writing (ρ=.24).

Table 49. Spearman's rho between C-test scores and TYS

	Turkish C-test (8-text)
TYS total	.65
TYS reading	.58
TYS listening	.67
TYS writing	**.24**
TYS speaking	.49
TYS level	.59

Note: all correlations statistically significant, $p < .001$ (except the bolded one, p<.005)

These correlations, except the ones with writing and speaking, are towards the higher end of the correlations stated in Eckes and Grotjahn (2006) where they did a survey study of the correlations between C-tests and standardized measures of language proficiency, which ranged between .33 and .87. The reading and listening correlations are also close to, although slightly lower than, the ones in Eckes (2014) where an 8-text C-test was used as a screening test for TestDaF (*r*=.61 to .73 for reading; .63 to .82 for listening). Regarding the main correlation of interest with TYS total score, it was within the range of other studies ranging between .55 and .87 (Eckes & Grotjahn, 2006) (see section 7.8.4 for a discussion).

7.7.4.2 Predictive power of C-test scores to estimate TYS performance (EQ 13)

To test this EQ, first ordinal logistic regression and then linear, quadratic and cubic regressions were conducted. For these regression analyses, the C-test total scores were taken as the predictor and four TYS levels as the dependent outcome.

Table 50 below shows the parameter estimates of the ordinal regression analysis.

Table 50. Parameter estimates of the ordinal regression analysis

	Estimate	*Std. Error*	*Wald*	*Sig*	*95% CI*
Threshold					
B2	*5.08*	*1.39*	*13.27*	*.000*	*[2.35, 7.81]*
C1	*7.45*	*1.55*	*23.16*	*.000*	*[4.41, 10.48]*
C2	*12.02*	*2.00*	*36.00*	*.000*	*[8.09, 15.95]*
Location					
C-test scores	*.08*	*.015*	*29.21*	*.000*	*[.52, .11]*

As seen in Table 50, there is a positive significant association between C-test total scores and TYS levels. For one-point increase in C-test total score, we would expect .08 increase in the ordered log odds of getting a higher TYS level. The significant Wald test result also indicated that the effect of C-test scores were significant.

Nevertheless, the proportional odds assumption of ordinal regression (one unit increase in the predictor brings the same predicted increase in outcome for all categories in the dependent variable) was violated since test of parallel lines provided could not be performed. This might possibly be due to the inequal number of participants in each TYS level (see Appendix 31 for test of parallel lines). Therefore, other regression models which would allow C-test scores to be entered in a more flexible way were explored. Hence, the curve estimation was conducted by comparing a linear, quadratic, and cubic predictor. Note that other ordinal models allowing for

polytomous C-test scores could be used instead of linear, quadratic and cubic models, and this should be explored in future research.

The results of the regression with linear, quadratic, and cubic models are shown on Table 51 below. They align with the scatterplots above showing that a quadratic or cubic model fits better than a linear model. C-test performance was able to predict 54.9% of the variance in TYS total scores in a quadratic model and 55.5% in a cubic model compared to its prediction of 52.1% of the variance in a linear model. Comparing cubic and quadratic models, they seem very similar to each other. Nevertheless, quadratic model was chosen over the cubic model since it showed a better fit as seen in the scatterplots above (Figure 30 and 31).

Table 51. Linear Regression Model Summary

Model	R	R^2	Adjusted R^2	Std Error of the Estimate
Linear	.726[a]	.527	.521	9.02
Quadratic	.749[a]	.561	.549	8.75
Cubic	.757[a]	.572	.555	8.69

Note: a. Predictors (constant): C-test total

Figure 32 shows the curve estimates for regressions when C-test scores are taken as linear, quadratic and cubic polynomials.

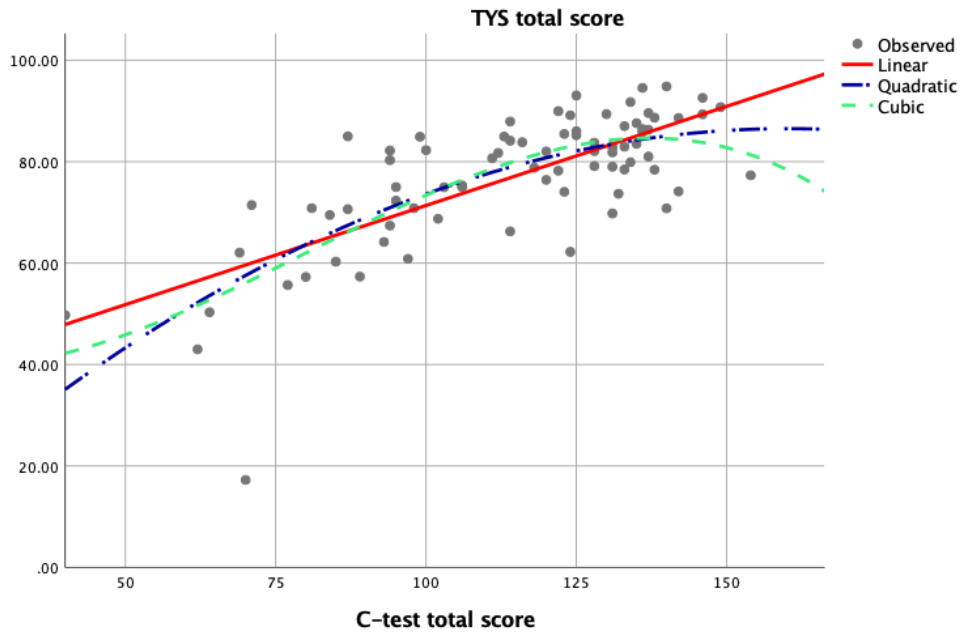


Figure 32. Curve Estimation

Given the slight non-linearity, even linear model would be adequate for these data. Nevertheless, quadratic and cubic model work better to distinguish lower level learners although they never reach the highest score on TYS and can only estimate TYS scores up until 135. Quadratic model shows there is a slight concave relationship between TYS score and C-test score whereby the change in TYS total score associated with a change in the C-test score reduces as we consider students with higher and higher C-test total scores. On the other hand, the deceleration of the cubic model is more obvious once students reach a score of 145 on the C-test, which means that predictions should not be made at the highest C-test scores on a cubic line. This supports the preference of a quadratic model over a cubic model (see Appendix 32 for the predicted TYS scores and levels by C-test scores using all regression models).

Following the curve estimation, the assumptions of normality and homoscedasticity (variance of residuals are constant across different values of the predictor variable) was checked based on the residuals of the preferred quadratic model. The histogram of standardized residual indicated that normality assumption

was met since it fit the bell-shaped normal distribution. Furthermore, Kolmogorov-Smirnov test of normality for residuals showed insignificant results (D (79) =.083, p=.200). Regarding homoscedasticity, the scatterplot of residuals against the standardized predicted TYS scores showed the constancy of residuals with equally distributed scores except one outlier (see Appendix 33 for standardized residual histogram and scatterplot).

7.7.4.3 Setting Cut Scores on C-test based on TYS (EQ 14)

To test EQ14, classification tables were created by cross tabulating the levels predicted by the C-test, which were derived from the regression models, and observed levels. Table 52 below shows the classification table created by using the levels predicted by the quadratic model (see Appendix 34 for the classification tables based on other regression models). It indicates how many students were placed into the correct TYS level based on the predicted TYS levels as well the percentage of correctly placed students in each level. It also shows the cut scores identified by the model where predicted TYS level shifts from one level to the next. These cut scores were calculated by cross tabulating TYS levels and C-test total scores and then counting the number of students predicted to be in each level (see Appendix 35 for classification table of C-test scores and TYS levels). For example, when the count of below B2 level candidates is shown as 3 as seen on the classification table below, the cut scores were determined based on the scores of the three lowest level candidates.

Table 52. Classification table for observed and predicted TYS levels by quadratic regression

Observed TYS levels		Predicted TYS levels			Total
		Below B2	B2	C1	
Cut scores		40-68	69-94	95-154	
Below B2	count	**3**	2	0	5
	%	**60%**	40%	0%	100%
B2	count	0	**7**	6	13

	%	0%	**53.8%**	46.2%	100%
C1	count	0	6	**44**	50
	%	0%	12%	**88%**	100%
C2	count	0	0	11	11
	%	0	0	100%	100%
Total	count	3	15	61	79
		3.8%	19%	77.2%	100%

Note: bolded numbers show the number and percentage of correctly placed students in each level.

As seen on Table 52, the C test was not effective in distinguishing among high levels, namely C1 and C2 levels, since all test takers who took C2 on TYS were identified as C1 according to their C-test scores. Overall, of the 79 students, the quadratic model placed 54 students (68%) correctly, while placing 8 students (10%) above their level (false positive), and 17 students (22%) below their level (false negative). Figure 33 below visualizes Table 52 by plotting the predicted TYS levels against C-test total scores while test takers are grouped according to their observed TYS levels.

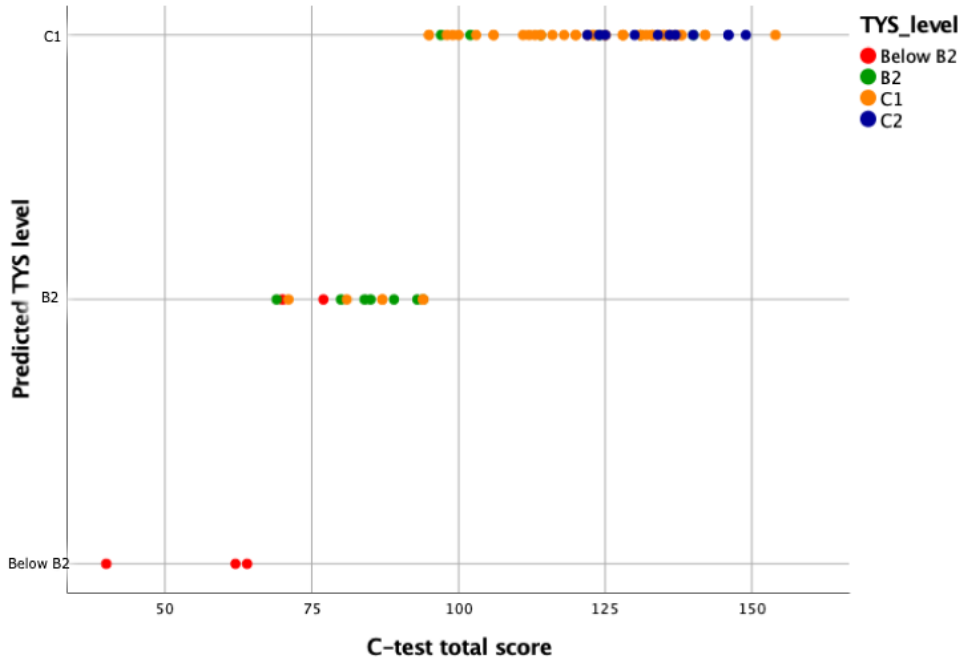


Figure 33. Scatterplot of predicted TYS levels against C-test total scores.

Figure 33 shows that the quadratic model was not able to predict any C2 levels, and thus, the highest level it can predict is C1 on the top row. All blue circles

(C2) are on this row since there is not a predicted C2 level. 88% of orange circles (C1) are on the top row showing that most of the C1 levels were correctly placed in the right level. More than half of the green circles (B2) were correctly placed in the second row while some were placed on the top row wrongly. Three out of five red circles (below B2) were placed correctly on the bottom row while the other two were wrongly categorised on the second row. Nevertheless, it is good that no successful TYS test taker (above B2 levels) was wrongly placed on the bottom row suggesting that they would fail the TYS.

Note that, classification tables were also created based on linear, cubic, and ordinal regression models. Among them, quadratic model provided the highest number of correctly placed students while also being more effective in separating between below B2 and B2 levels as well between B2 and C1 levels (see Appendix 34).

7.7.5 Results for the Decision Inference

7.7.5.1 Perceptions of Stakeholders (EQ15)

This section relates to EQ 15 *"what are the perceptions of stakeholders involving instructors of Turkish and TYS candidates regarding the usefulness of the Turkish C-test to predict TYS levels and practice Turkish?".*

7.7.5.1.1 Instructors of Turkish

Initially, a series of 5-point Likert-scale questions elicited instructors' (N=34) opinions regarding TYS about the: 1) appropriateness and fairness of the TYS to estimate general Turkish L2 ability, and 2) usefulness of the TYS to determine international students' admissions to Turkish-medium universities. Following this, they were asked 5-point Likert-scale questions about the Turkish C-test regarding the: 1) clarity of the Turkish C-test example, 2) sufficiency of the Turkish C-test

instructions, 3) appropriateness and fairness of the Turkish C-test to estimate Turkish L2 (general) ability 4) usefulness of the Turkish C-test to quickly estimate whether candidates are ready to take TYS, and 5) usefulness of the Turkish C-test to estimate TYS levels (see Appendix 21 for instructor survey)

Instructors were mainly positive about the appropriateness, fairness, and usefulness of TYS. Although the EQ 15 did not directly address TYS, eliciting instructors' views of TYS was useful to justify the face validity of the TYS as a criterion test. Table 53 below shows Likert-scale statements and the percentage of instructors' agreement/disagreement with these statements.

Table 53. Instructors' perception of the TYS

	Strongly or somewhat agree	Neither agree nor disagree	Strongly or somewhat disagree
TYS is a good and fair estimate of Turkish language ability.	88%	9%	3%
TYS is useful to determine international students" admission to Turkish-medium universities	95%	5%	10%

Regarding the clarity and sufficiency of the Turkish C-test example and instructions, instructors were very positive as seen on Table 54 below.

Table 54. Instructors' perception of the Turkish C-test instructions and example

	Strongly or somewhat agree	Neither agree nor disagree	Strongly or somewhat disagree
The Turkish C-test example was clear	100%		
The Turkish C-test instructions provided enough information about the test format	94%	6%	

In relation to the C-test, Table 55 below shows instructors' responses to various Likert scale questions.

Table 55. Instructors' perception of the Turkish C-test

	Strongly or somewhat agree	Neither agree nor disagree	Strongly or somewhat disagree
The Turkish C-test above is a good and fair estimate of Turkish language ability	47%	29%	24%
The Turkish C-test above will be useful to quickly estimate whether students are ready to take TYS (i.e., whether students are at least B2 level)"	47%	9%	44%
The Turkish C-test above will be useful to estimate student levels attained by TYS (below B2, B2, C1, C2)"	39%	53%	9%

It is difficult to observe a common pattern about instructors' responses to these statements. To understand why the instructors were sceptical or not so positive about the C-test, thematic analysis was conducted on the open-ended survey questions asking to elaborate on Likert-scale survey questions and two main interview questions "*What is your impression of the Turkish C-test?*" and *"Do you think Turkish C-test can be used as predictive of student performance on TYS? Why or why not?"*. Based on the thematic analysis, one theme in relation to the usefulness of the Turkish C-test to predict TYS levels was generated: insufficiency to measure language skills. Note that interviews were conducted with only two instructors while there were 34 instructors participated in the survey, so results should be interpreted without overgeneralization.

Theme: Insufficiency to Measure Language Skills

Instructors considered the C-test insufficient to measure language skills based on the reasons summarised on Figure 34.

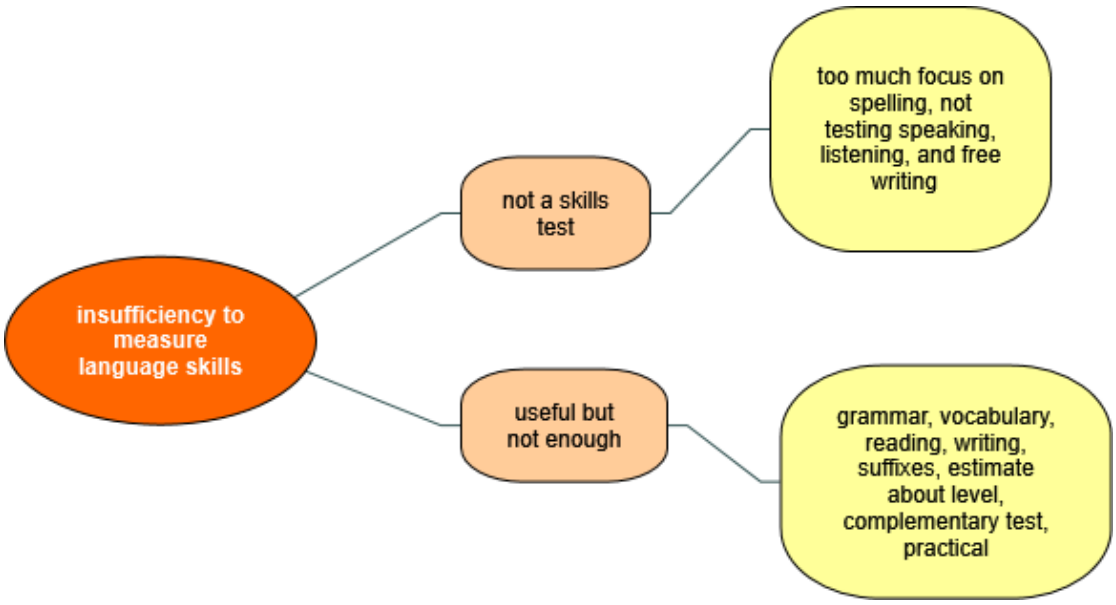


Figure 34. Theme 1 Insufficiency to measure language skills

Thirteen teachers out of thirty-four considered the C-test as insufficient since it was not a skill test involving speaking, listening and free writing. Teacher 17 implied the lack of oral skills in C-test when he compared it with TYS which has four language skill sections. Thus, he considered C-test would not be enough to determine learners' levels since it is "based on grammar rules and writing skill".

> TYS is an exam that involves four main language skills. Therefore, it is not enough to determine a student's language level based on grammar rules and writing skill. Speaking and listening are an important part of TYS, and students have the most difficulty in listening section.

Quantitative findings contradicted this comment showing that C-test scores had the highest correlation with the TYS listening section (ρ=.67). Furthermore, candidates were found to have the lowest scores in TYS writing rather than listening by having a mean score of 17.75 out of 25.

Teacher 4 mentioned that the focus on spelling in calculating C-test scores is not fair since every learner may not be good at filling the blanks when she was asked whether C-test is a good and fair estimate of Turkish language ability. She also added that C-test lacks oral components, so, it cannot be used for university entrance.

> I believe there is more to language ability than being able to complete words in a paragraph. Spelling being able to affect your score so much is also a reason why I don't think it's fair. Even though a student is proficient and speaks Turkish in their social lives or reads books in Turkish, not everyone would be able to complete the words. Also, if this is used for university entrance, how will we know if students have the ability to listen and understand the lectures or not?

However, it was not the claim of this study to use C-test as a university entrance exam or a replacement exam for TYS, and test users should be careful about the claims they can make based on the scores. Rather, the main aim of the study was to investigate candidate readiness for TYS (i.e., whether they would pass or fail TYS) while the sub-aim was to predict all TYS levels since different institutions have different minimum requirements from students.

Seven teachers thought that the C-test is more useful in testing grammar, vocabulary, reading, and writing as exemplified on the excerpts below.

Teacher 6

> This type of test is only useful for determining learners' ability in grammar, vocabulary, and maybe a little reading. But, as we saw the overview of TYS, it consists of 4 major skills. So, we cannot say this type of test is a fair way to estimate a student's overall language proficiency

Teacher 31

> I believe C-test is to assess language skill rather than language proficiency. C-test can only address two language skills (reading and writing), however, TYS involves four language skills. Also, C-test is based on only reading and filling the gaps. Therefore, a candidate who gets a high score on C-test might be unsuccessful at TYS or vice versa. Overall, these two exams are too different from each other to be related.

Again, in contrast to teachers' comments, C-test had the highest correlation with TYS listening (ρ=.67) and total scores (ρ=.65) rather than reading (ρ=.58) and writing (.24) scores. Regression analyses were also positive in showing that C-test and TYS scores were related. Quadratic model fit the data the most in that C-test scores can predict the candidate performance on TYS until they reach C2 level. Furthermore,

it was able to classify 68% of students in the correct TYS levels based on C-test scores. The percentage of correct classification was 60% in below B2 level, %53.8 in B2 level, and 88% in C1 level while the C-test placed no candidates in C2 level. There was no occasion when the C-test predicted a successful TSY candidate (B2 and above) as unsuccessful (below B2) contrasting Teacher 31's comments.

Two teachers were positive in that the C-test would be useful to test a learner's proficiency and see whether they are ready to take TYS since Turkish is an agglutinative language (see section 3.3) and thus, being able to complete suffixes is important.

> Mastering the word endings in Turkish is a clear indication of student's proficiency. C-tests above seem very difficult, but it is a good way to test a learner's proficiency level.

> We see that most of our students, especially the ones below B2 level (A1, A2, B1), cannot write words and suffixes correctly. With this test, we can identify them and guess whether they are ready to take TYS.

Given the argued insufficiencies of the C-test to measure all language skills, two other teachers suggested to use it as a complementary test with other exams, in particular speaking and listening tests.

> I think a student who is unsuccessful at C-test might be not unsuccessful in listening and speaking skills. I also think C-test is not enough in measuring writing skill as well since writing is only based on filling in the gaps and does not involve creative writing or structured writing. But I think it would be more appropriate to use C-test as a complementary instrument with other exams

> I think C-tests can be useful to estimate the learners' proficiency levels, but I think it is not enough to determine their exact level. Many researchers suggest using open-ended questions, maybe integrated tasks and also, nowadays very popular, reading into writing tests. Just in my opinion, I think we should use all of them together to estimate the exact level.

The C-test can be used as complementary with other tests as suggested by teachers depending on test purpose. For example, a German C-test is used alongside a reading comprehension and listening comprehension test for placing university

students into the correct language classrooms aligned with a four-year curriculum in the German FL department of a US university (Norris, 2006). However, since the Turkish C-test is evaluated as a short and quick low-stakes screening test in this study, it was not considered to combine the C-test with other tests. A teacher pointed out to this use of the C-test when asked in which contexts C-test would function better.

> Since it is easy to administer, maybe for diagnostic purposes. But I am not sure about placement purposes. Because for like placement purposes, I would probably want more comprehensive test like TYS maybe. But for diagnostic purposes, yeah it can be a good and easy test, I guess.... It depends on the purpose. So, if I am gonna do a placement test, probably I would just want them to take TYS. But for other like, let's say for exam purposes, I may ask them to take the C-test. I don't know you know like how well the C-test is gonna measure their proficiency level compared to TYS. If the levels are gonna be similar, then I would definitely ask them to use C-test. Why not use the easy and free one?

Again, two other teachers pointed out to the practicality of the C-test in terms of time and costs. However, they had doubts about which levels such a quick test can distinguish.

> Teacher 4
>
> I do believe that the test is time and cost effective. The test can tell you if a student is higher than a certain level. But the questions of how high and what level cannot be answered.
>
> Teacher10
>
> Regarding the second statement "The Turkish C-test will be useful to quickly estimate whether students are ready to take TYS", I strongly agree because this test might be useful in terms of practical and quick use. I am indecisive about the third question about C-test predicting TYS levels because in order to predict the differences between levels, the exam scales with a wider variety might be required.

In contrast to teacher 4's beliefs, regression analyses and classification tables established what levels the C-test can distinguish. The C-test was able to distinguish learners below B2, B2, and C1 level. However, it was not able to separate C1 learners from C2 learners. Therefore, to distinguish among high level learners, the Turkish C-

test can be used together with other tests or the easier texts might be replaced with more difficult ones.

Overall, instructors were not positive about the usefulness of the Turkish C-test to predict TYS levels since it did not include skill sections, in particular listening and speaking, in contrast to TYS. This view contradicted the quantitative findings which showed the C-test was able to place 68% of the candidates in the right level and had the highest correlation with TYS listening.

7.7.5.1.2 TYS Candidates

Based on the thematic analysis of the survey and interview responses of TYS candidates, the following two themes in relation to the EQ15 "perceptions of TYS candidates regarding the usefulness of the Turkish C-test to predict TYS levels and practice Turkish": (1) candidates' understanding of their need to learn more; (2) difference of format with TYS.

Theme 1: Candidates' understanding of their need to learn more

The first theme was related to what test takers thought about their performance on the Turkish C-test and what impacts the test had on them as summarised in Figure 35. It was generated from interview responses (N=13) since the feedback survey did not include any open-ended questions about the potential impacts of the Turkish C-test. Candidates' responses were contrasting instructors' opinions in that teachers were more conservative towards the C-test while candidates were positive acknowledging its potential benefits on themselves. Note that while qualitative data were collected mostly through surveys from teachers (N=34), with only two instructor interviews, thirteen interviews were conducted with candidates in addition to candidate surveys (N=76).

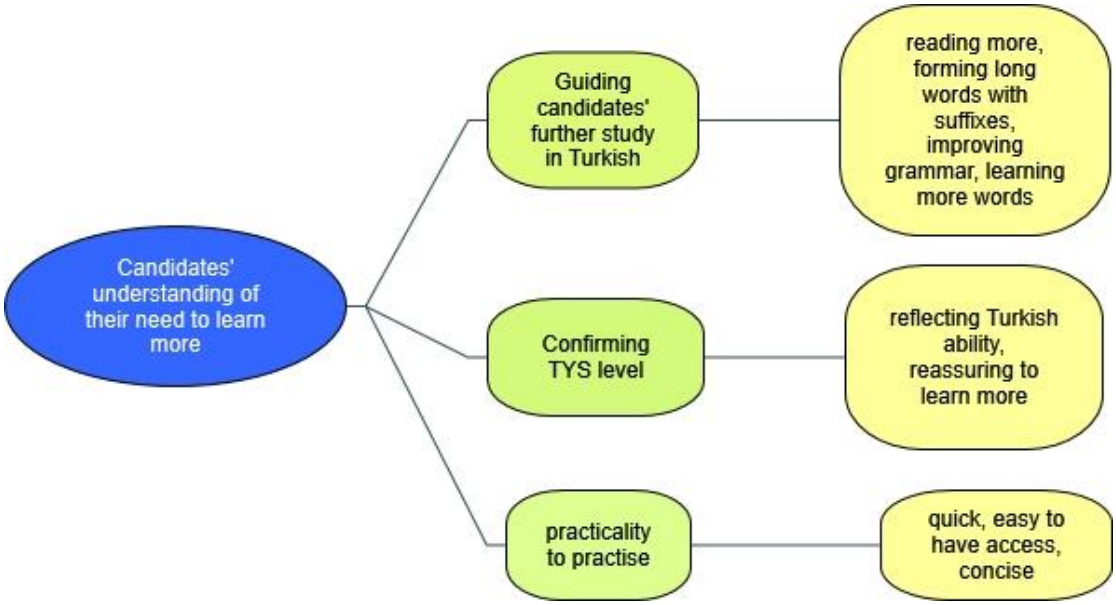


Figure 35. Theme 1 Candidates' understanding of their need to learn more

Eight out of thirteen interviewees reported that the test showed there is more they need to learn. Phrases such as "motivation to learn more", "encouraging to learn new words", and "positive effect" kept recurring across interviewees. As seen below, Examinee 8 said that the C-test was motivating for her to read more books since she realised there was more to learn. Her TYS level was C2 and she got 125 (out of 160) on the C-test. She seems to have found the C-test encouraging to continue learning Turkish even after getting the highest level on TYS.

> After I took TYS, I was thinking OK I know Turkish very well. But now I realise I should read more books because my score on C-test was lower, which means there are still some points I don't know about. After taking C-test, I decided to read more books. It was a motivation for me to learn new things.

Other examinees said that C-test helped them to realise their weaknesses in Turkish, such as suffixes and reading academic topics. For example, Examinee 35 (TYS C1 level), who was doing an MA in a Turkish-medium university in Turkey, said that he had difficulty in understanding academic terms in Turkish, and this was also reflected in his performance on the C-test.

> Since I use Azerbaijani words a lot when I speak, professors sometimes don't understand me. I also have difficulty in understanding academic terms in

> Turkish... With this test, students can see at which points they are having difficulty. For example, I realised I have more difficulty in academic texts, and I should read more academic texts.

Similarly, Examinee 10 said that the C-test showed her weakness in Turkish, which is suffixes, as seen on Excerpt 50 . This was a wake-up call for her after getting C1 on TYS.

> My confidence in my Turkish was boosted by TYS result and I got the wake-up call. But it is a positive effect because I don't think one really finishes learning a language. For example, I started learning English when I was 8. Now, I am 32, and I am still learning it. It is a process for all the life. This test showed me how I should work in my further Turkish study. It showed my weakness very clearly. My weakness is forming the long words with suffixes.

Regarding the same examinee, upon being asked whether she would use this test to practice before exam preparation, she strongly agreed, which addressed the EQ 15. However, she added that if not for exam preparation and beginner level studies, she would not use it since she prefers to learn language through more authentic ways.

> Yes, definitely, and I would prepare accordingly. But, if I was not preparing for the exam, I wouldn't. I would find it quite frustrating given that I have a problem with suffixes. I only study grammar tests when I am at the very beginning of learning a language. Otherwise, I don't think it is really efficient. I think the most effective way of learning a language is books, movies, conversation with the native speakers. And it is also more fun. I don't like studying textbooks.

The comment by Examinee 10 above corresponded with the quantitative findings that the C-test is more effective for levels under C2 since it could not distinguish C2 from C1 level. At C2 level, learners might benefit better from more authentic ways of practicing language.

Two other participants commented that their performance on the C-test reflected their TYS level. As seen below, Examinee 1 said that C-test was 'reassuring' in the sense that he struggled with the last texts since he got C1 on TYS and there is more to learn. Interestingly, he also mentioned that he knew when he got wrong and it was an overall positive experience for him.

> Most of the times when I got things wrong, I knew that I got it wrong. I was sort of like I really don't know what this is supposed to be. I am just gonna fill it like what sounds the best to me. Even on the tougher texts, I was still getting answers correct so I felt like positive at least and most of the things that I got wrong, I sort of knew that I was having issues. I wasn't also sure whether there would be any left blank entirely. Overall, I felt good because the ones that I really struggled with were the last two or three. So, I thought it felt pretty accurate to me. I was like oh yeah, I made small mistakes at the beginning and obviously kind of increases throughout. And by the end, I was like sitting there, scratching my head. So, I thought it seemed reasonable to me. Each text was like getting harder. Since I got C1 on TYS, there is still a whole level where I should be. So, it is reassuring to be like there is more I need to learn.

Similarly, Examinee 69 (TYS C1 level) said he performed the same in both TYS and C-test as seen below. Qualitative findings confirmed that the C-test placed him in TYS C1 level. He added that the C-test confirmed his difficulty with reading and revealed his problem with words which doesn't agree to vowel harmony rules.

> I think I did approximately the same in both tests. I think the texts of the C-test reflects my ability carefully, but I don't know why... One of the things that this test confirmed is that I need to read more. The main difficulty with Turkish I am facing right now is reading. When a sentence is long, my brain cannot understand it automatically. If a sentence takes several lines, then I have to analyse where is the subject, where is the verb, how are the words connected. I have to stop and think. I think the only way to increase this ability is to read a lot. That is what I am trying to do right now. With C-test, I also realised I don't know some words where the vowel harmony is broken.

The practicality of the C-test to practice Turkish before taking TYS also came up. Words such as "quick", "easy to access", and "concise" kept recurring. As seen below, Examinee 1 said that he was looking for such a quick test to estimate TYS levels before taking the exam since the stakes were high on TYS. However, he could not find such a test and this study seemed to "fill a needed gap" in Turkish education according to him.

> TYS only gives a level over B2. For me, it was a bit nerve wracking. I assumed I was gonna get at least B2, but there was always the possibility that I could mess up the exam and get something lower and then just not even receive a score. So, basically, I would have wasted whatever 200 TL with nothing to show for it. So, I think yes. Obviously, it doesn't do speaking listening, but again still there is some correlation between being advanced

> level and being able to speak and listen. I thought it was a fair assessment. That's also pretty quick and easy especially when you compare it to the exam. It was actually something that I was looking for before TYS. But, as far as I could find, that sort of thing doesn't exist on the internet at least. So, I think what you created would definitely help to fill a needed gap for at least Turkish education. This works for what it purports to be which is something quick and you can have access easily. Fairly accurate indicator of where you can be on TYS. What it does is quite good.

Similarly, Examinee 78, who was unsuccessful on TYS, also said that the C-test would be good to practise before TYS due to its concise structure.

> I think the Turkish C-test can be a good assessment before taking TYS. It is very concise and easy to understand the structure as the sentences are not so long.

Overall, theme 1 showed that candidates were very positive in using the Turkish C-test to practise their Turkish skills before taking TYS due to the following reasons: (1) Turkish C-test showed them what they need to practise more (i.e., reading more academic texts, forming long words with suffixes); (2) Turkish C-test is quick and easy to have access.

Theme 2: Difference in format between the C-test and the TYS

The second theme was related to the C-test having a different format from TYS as summarised in Figure 36 below. These format differences were divided into the following three sub-themes which will be explained in this section: 1) lack of oral and free writing component, 2) candidates' unfamiliarity with the C-test format, 3) assessing overall ability while not assessing four main skills individually. The first and the third subthemes overlap with instructors' perception of the C-test.

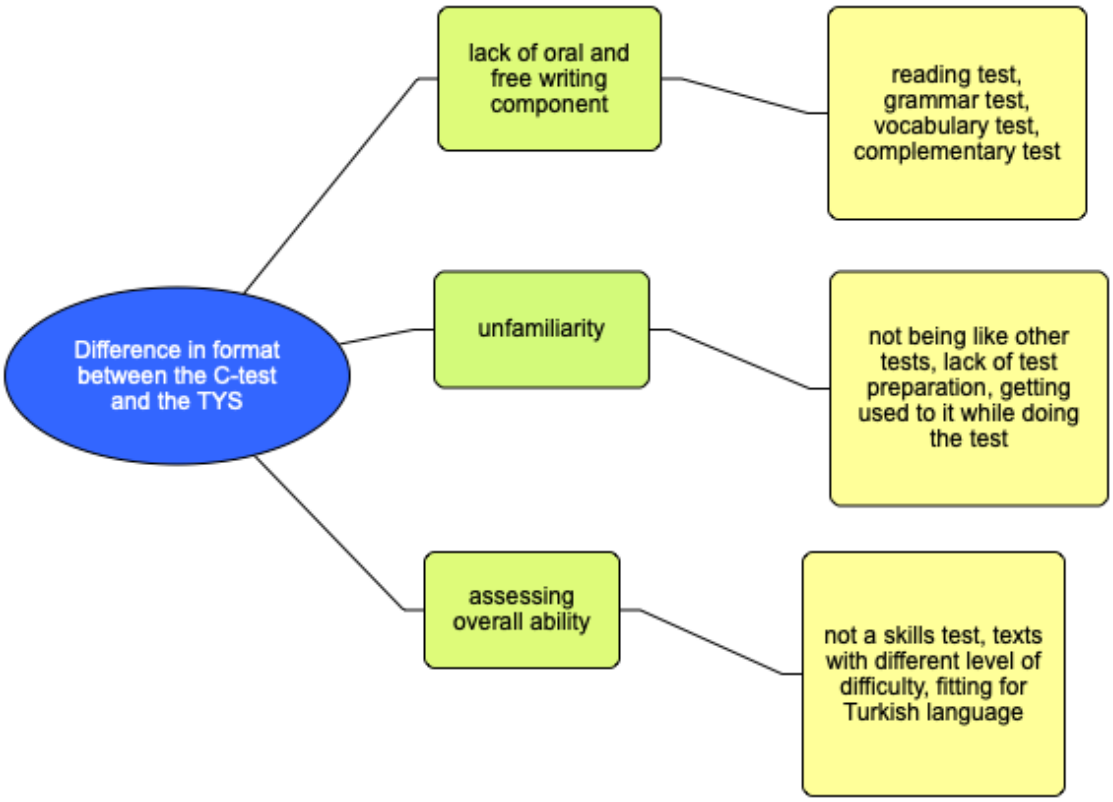


Figure 36. Theme 2: Difference in format between the C-test and the TYS

Thirty percent of the participants (24 out of 78) mentioned that the C-test did not have any oral component while the TYS had speaking and listening sections. Also, twenty four percent of the participants commented on that the C-test did not involve free writing as in the TYS. As seen below, Examinee 1 (TYS C1 level) said that C-test can be useful to practice for reading and writing and even so, the format of reading and writing on TYS was different than it was in C-test.

> TYS has reading, writing, speaking, and listening. So, it has four skills and I think that is important. You presented your Turkish examination and since it is online, you can only do reading and writing. But having speaking and listening sections is also very important to holistically evaluate somebody's language. As for practicing before an exam, it could be useful to check reading/writing skills, although the way that proficiency test proceeded was more multiple choice for reading and then free writing.

Similarly, Examinee 7 (TYS C1 level) commented that the C-test was good, but not enough since he was worried that the C-test would not show "a complete

picture" of his language skills as seen below. Therefore, he suggested improving the test by adding, for instance, listening, speaking and writing sections.

> Yes, it's really a good test, but it's not enough. And maybe it's just part of it, maybe there's still a big part of the test. This is to the fact that if only on this part of the judge, then I'm afraid that it will not display a complete picture of language knowledge. Perhaps it needs to be improved, add Listening, Speaking, Writing etc. I do not know, maybe I'm reasoning classically, but this is only my subjective opinion. In general, I think the test is moving in the right direction, in order to master the language, you need to know the spelling well, that is what C-test is aiming.

As previously explained while reporting instructors' perception (section 7.7.5.1.1), using the C-test as complementary to other tests is not feasible for the screening purposes of this study. According to the regression and IRT analyses, it is true that C-test could not show a "complete picture" of language skills for high level examinees (over C1 level). However, note that examinees over C1 level are not at-risk (i.e., low level) learners and would not need a screening test to see whether they are ready to take TYS.

Nine participants considered the C-test as a grammar test. As exemplified below, Examinee 25 (C1 level) also suggested to use the C-test alongside other measures in order to predict TYS levels since she considered the C-test as a grammar and spelling test. Upon being asked whether she would use the C-test to practice before TYS, she said that she would use it to practice grammar.

> I don't see how it assesses anyone's overall language proficiency without a speaking part or a free writing part. I think it can assess someone's proficiency with grammar and spelling, and even with that, I think still limited with grammar because some people do better on multiple-choice tests. Just a test in the format you provided, I don't know how accurately it measures someone's level. I think it can roughly predict, but it should be combined with other measures as well.

> Yes, I think it would be good to practice grammar and I would like to practice grammar before TYS.

Similarly, Examinee 30 (TYS C1 level) suggested to include the C-test on the TYS since the C-test "seemed to involve everything except listening and speaking" as seen below.

> My Turkish is good if I look at TYS result, but according to C-test result, my Turkish is not that good. It seemed to me that the two tests measure different things. It would be good if Turkish C-test was included in TYS because C-test seemed involve everything except listening and speaking.

Interestingly, she thought her Turkish level was not good according to the C-test although she got 120 points from the C-test and her TYS level was C1 with 76.41 total score. Thus, she was placed in the correct C1 level according to her C-test result since scores above 94 were placed in C1 level using quadratic regression as explained in section 7.7.4.3. This examinee's perception of her C-test performance "as not that good" shows the discrepancy of performance expectations between test takers and test developers, which aligns with the C-test literature (Sigott, 2004) Clearly, there was no expectation from test takers to do 20 out of 20 in all texts to be successful in the C-test since the test was norm-referenced. Examinee 31 added that the reason why she found the test difficult was her unfamiliarity with the C-test format because she was used to multiple-choice tests.

> Actually, it was a bit hard for me because it was a different exam type. There were no multiple-choice questions as a clue. There were only letters at the beginning of words, and I found the end of the words by guessing. I am not used to this type of a test since it is not multiple choice.

Unfamiliarity with the C-test format came up among seven other examinees. While Examinee 31 knew what he would expect on TYS before taking the exam and made necessary test preparations, he had no clue about what he would see before taking the C-test and did not have test preparation. While he got C1 level with a total score of 85 on TYS, he got a score of 87 on the C-test and was placed in B2 level.

Thus, the C-test placed him in lower level than his actual level, which he attributed to his lack of exam preparation.

> I got a good result from TYS and I feel really proud of me. But when I took your test, it was really difficult. Maybe because I had no idea what the test was about. For TYS, I studied, and I prepared myself to do it and I was psychologically prepared to do it. I knew that I was facing an exam. In your exam, I had no real idea that I would do this kind of a test. I haven't had that previous preparation. That was different, I think. Also, when I was doing the test, I thought that maybe the test will be difficult even to native Turkish speakers.

Examinee 1 (TYS C1 level) mentioned that although the test was not like any other online tests he had seen earlier and he found the test 'scary' at the beginning, he got used to the test during the process of doing it. He said that the ordering of the texts from the easiest to the most difficult made the process of familiarization with the test easier.

> Your exam was kind of scary at first. I was like oh man you know because it sort of called back other online exams that sort of filling in the blank tests that give you list of choices, which obviously is a little bit easier, right. Obviously, you get over that initial shock and then you are like OK I could do that, that's pretty easy, especially the first page or two are fairly basic. So, I thought at least it was like a good way to start. That made a lot of sense to figure out how you should be conceptualizing about it

Nevertheless, unfamiliarity with the C-test format did not seem to create frustration or disappointment on the side of the examinees. Examinees found the test "eye-opening", "optional" or "interesting" as exemplified by Examinee 10 (TYS C1 level) below. However, note that there was no interviewee below B2 level in this study.

> My score was quite low sixty something percent correct. It was difficult, but also very eye-opening. It is original. Because never before I saw such a test and while preparing for TYS, I tried most of the online tests and also books. But I didn't really find tests which worked on this part of the suffixes. I would love if I could find study material that would help me with the suffix problem. I would like to find other similar tests because this is my weak point and I would like to work in this direction.

There was only one examinee (Examinee 69) who said that he was familiar with the C-test format, and this examinee was working as a software developer (computer engineer). He said that he would prefer C-test to multiple-choice test since it is easier to guess on multiple-choice tests. He got a score of 101 on C-test and his TYS level was C1. He was the highest scoring examinee among all interviewees, not all examinees though.

> It wasn't something new for me because I know this type of test, C stands for cloze. And I know this technique of learning a language, the cloze deletion test. It is used for example in some spaced repetition applications. But I never saw it being used as a testing technique. It is an interesting approach. I liked it. I prefer this to multiple choice because in multiple-choice, you can guess. But in this test, you have to know. As a self-assessment, it is very good.

Despite most examinees' unfamiliarity with the C-test format and their perception of the C-test as lacking oral and free writing components, six examinees said that the C-test would assess overall language ability even if not each language skill individually. Examinee 1 implied that there is a correlation between doing well on the C-test, in particular on the later texts, and oral skills to some extent. Correlational analyses supported this view that Turkish C-test had the highest correlation with the TYS listening section.

> If I go and do one of the ones that are multiple choice, if I can recognize the correct answer, then I can just plug it in. It is much harder actually to produce the correct answer yourself even if you sort of know what it is supposed to be like whether it is dative case or accusative case, or you are supposed to use the plural or not. That can kind of be a lot of things to keep in mind when you are actually filling out the answers. I think overall it will be very challenging to be able to do very well in the later sections without having an overall positive ability in Turkish. I think just by nature learning going through academic materials and taking classes you take up those skills as well it is almost unreasonable to say that somebody could get all the answers correct on that last section, but not able to say a single word. So, maybe it is not a hundred percent accurate in each skill area, but overall it does accurately assess where you are in this continuum of very basic to very advanced.

One reason why the examinees considered that C-test assesses overall language ability was due to the differing level of difficulty, in particularly, on the last

two texts. Even C1 level learners found the words on the last two texts difficult.

Examinee 1 (TYS C1 level) said that he was stuck on these last two texts.

> Then as it went further on, I think the first five or maybe six text, I knew all the words that I should be using. I didn't see a blank and then be like I have no idea what goes here. I always knew maybe it's like this kind of ending or maybe not sure it might be this. But in seventh and eighth text, I had no idea what I could even put here. Nothing was coming to my mind.

Along a similar line, Examinee 69 (TYS C1 level) said that while he was filling out the gaps even without "thinking" on the first several texts, he had to think and "reread several times" on the last two texts. He got 19 points out of 20 on the last and most difficult text.

> Apparently, you chose the texts to be in ascending difficulty order, from the easiest to the most difficult. With the most difficult one, I had a problem, I had to think a lot. For the first several texts, I didn't even think. I could easily guess the word. But for the last two ones, I had to reread several times.

Similarly, Examinee 8, who got C2 on TYS, said that she got more mistakes on the last text since she filled in the gaps without paying attention to the whole sentence and context. She got 14 points on the last text.

> The last texts were a bit more difficult than TYS. On C-test, words did not have their last parts. I got more mistakes on the last text since I sometimes wrote the answers without considering the whole sentence and other sentences. I got confused about whether it is plural or singular and which tense I should use. I realised I should have read the other sentences while evaluating one sentence. The last text also had an abstract topic, which I could not exactly understand

On the other hand, Examinee 38, who got B2 on TYS and was classified as B2 on the C-test, said that he guessed the blanks without understanding the meanings even on the most difficult texts.

> The former part was somewhat easy for me, but the later part was difficult because it required lots of vocabulary. Vocabulary was higher level than my Turkish vocabulary. Especially in the later part of the questions, I couldn't comprehend the whole meaning of the sentences, but I could guess the blank or the type of the word from the previous and following words. I thought it doesn't require me to understand all the sentences. I didn't have to understand all the sentences.

Therefore, it seemed that higher level learners seemed to try more on the later texts by rereading a few times and trying to understand the whole meaning of the paragraph while lower levels thought they could guess the meaning of the words without reading even on the most difficult texts. This aligned with the “fluid” structure of the C-test (Sigott, 2004) since learners with different proficiency levels may use different amount of context solve a C-test item.

Another reason why some examinees considered that the C-test assesses overall language ability was due to that C-test was fitting for the Turkish language due to its agglutinative structure as exemplified in the two excerpts below. Thus, the test would help them to learn suffixes, which is a major challenge in Turkish.

> I agree that this test is good for estimating the Turkish language ability, given that Turkish is an agglutinative language, the correct use of suffixes being key. It was very fitting for Turkish because it was testing exactly the most difficult part of the language at least for me.

> I think people who learn Turkish make mistakes about what endings words should have. It is a test that makes you think a lot from this perspective. You cannot write the answer very easily. You say that I need to think a bit here. I never took such a test where I complete the letters before. But, I think it is really useful. it was very interesting to stretch my brain. In general, I liked the test, I think it affects the most "problem" moments in the study of Turkish and the correctness of writing words, with endings.

Overall, theme 2 showed that candidates view the C-test scores as predictive of TYS results to some extent, which aligned with the quantitative findings that C-test scores could predict TYS levels up until C2 level. Therefore, some of the candidates suggested to use the C-test as complementary to other tests, in particular, listening and speaking since it lacks oral section. This point was also made by instructors previously. However, given the screening purpose of the C-test in this study, there was no need to make the test duration longer.

7.8 Discussion

This section discusses the screening and predictive potential of the Turkish C-test for the TYS based on a priori validity evidence reported in Chapter 5 (i.e., test development decisions) and a posteriori validity evidence reported in the current Chapter 7 (i.e., setting cut scores on the C-test for each TYS level). In order to do so, it answers to what extent each assumption was accepted or rejected under each relevant inference of the interpretive argument.

7.8.1 Theoretical Grounds

The evidence collected for the assumptions of the theoretical grounds is based on the literature and its discussion is the same as in Chapter 6 except the inclusion of the TYS construct (see section 6.8.1). Assumption 1 "the common core of general language proficiency is inclusive of, but not limited to, grammar and lexis" was accepted based on the analysis of the L2 proficiency models described in section 2.2.1. Although there was not a consensus L2 proficiency model, all models agreed that the general components of L2 proficiency are grammar and lexis. Thus, this study defined the general language proficiency as a 'unitary' concept (Oller, 1971) involving but not limited to these general elements of language proficiency.

Assumption 2 "C-test can assess general language proficiency" was accepted based on a considerable amount of literature (see Grotjahn, 2017 for the latest C-test bibliography). C-test was chosen as the measurement method in this study since it is a short-cut estimate of general language proficiency based on the reduced redundancy principle (i.e., Eckes, Grotjahn, 2006; Norris, 2018; Sigott, 2004). C-tests are feasible for screening purposes since they require knowledge of grammar and lexis in an embedded context and can be completed in a short amount of time. Assumption 3 "TYS is a standardised test of Turkish language proficiency aligned with CEFR" was

accepted based on the exam description (see section 3.2.2). Although there is no publicly available study about the validation of the TYS, it is a standardised exam recognised at the educational and governmental level both in and outside of Turkey. It involves reading, writing, listening, speaking sections and ranges between B2 and C2 CEFR levels. Similar to TYS, Turkish C-test includes a range of texts with different difficulty levels and these texts relate to both academic and general topics aligning with the texts of the TYS.

7.8.2 Scoring

The evidence collected for the assumptions of the scoring inference is based on the decisions and analyses done during the test development stage (see Chapter 5) as well as analyses conducted in the current chapter.

Assumption 4 'text selection and word deletion procedures are appropriate to cover a range of L2 learners in terms of Turkish language abilities' was accepted looking at the expert judgement and teacher perception of the text difficulty. Similar to Lee-Ellis (2006) who created a Korean C-test ranging between ILR levels of 1 and 2+, the 8 texts of the Turkish C-test ranged between ILR 1 and 3+. It would be ideal to rate the texts according to the CEFR levels as well since TYS is CEFR aligned. However, given that texts were first piloted with US college students (during test development) and there were no funds to recruit Turkish instructors familiar with CEFR to rate the texts, texts were only rated according to the ILR levels, one of the two major frameworks used in the USA. The number of different levels the Turkish C-test was able to distinguish was also aligned with Norris (2006) where a German C-test was developed to place college students across four years of a German curriculum. The small group of high-level learners who weren't covered by the most difficult text was not considered an issue since the main aim of the Turkish C-test was

to identify lower-level learners at risk of failing the TYS or not meeting the minimum requirements required by the institutions. If future researchers needed a Turkish C-test to distinguish among higher level learners, they should replace the easier texts (i.e., Text 1) with more difficult ones.

Assumption 5 about the fitness of items was met as evidenced by the acceptable item fit statistics of all the texts. Furthermore, all texts had acceptable point-biserial values over .80 except the easiest texts (T1, T3, T4). The reason why easier texts had lower point-biserial values is probably because only 5 TYS candidates who failed the TYS (below B2 level) participated in this study. It would be ideal to have equal number of participants from each TYS level. However, it was challenging to motivate lower ability learners to participate in the study.

Assumptions 6 and 7 about the appropriateness and consistency of the scoring criteria were also met by the automatic scoring and the answer key involving undeleted versions of the words as well as alternative answers that emerged during the piloting. Online administration of the test (through the Learnclick platform) would enable future test administrators to benefit from automatic scoring and the trialled answer key so they would not have to spend time to mark the C-tests and develop a new answer key.

7.8.3 Generalization

The evidence regarding the generalization inference was based on the reliability analysis conducted in this chapter and the literature review. Assumption 8 about the internal consistency of the C-test texts was satisfied by a high reliability value which was at the higher end of the Cronbach alpha values stated in Eckes and Grotjahn (2006) ranging between .75 and .96 (see Table 1 in 2.3.2.3). As discussed in section 6.8.3, high reliability is one of the key features of C-tests, which results from the

consistency in the design of each C-test text (Roever, 2018). The reliability of the 8-text C-test in this chapter was slightly higher than the reliability of the 6-text C-test in study 1 probably due to the larger number of texts, but there was not a discernible difference. Assumption 9 about the sufficiency of the sample size was partially met based on the literature review on the validation studies in other LCTLs. As explained in section 6.8.3, at least 210 participants would be ideal given that 10 observations would be required per category for polytomous C-test scores. However, it was very challenging to reach such a large sample size. 79 TYS candidates were reached out of 3,477 candidates who took TYS in 2018 or 2019 January. More candidates could have been reached if there was no recruitment condition for candidates who have taken the TYS within the last one year. However, then, there would be a higher risk of whether candidates' proficiency levels have changed since they took the test.

7.8.4 Extrapolation

The evidence for the extrapolation inference was based on the correlational and regression analyses conducted in this chapter.

Assumption 10 about the correlation between C-test scores and self-perceived proficiency was accepted as indicated by the positive and significant correlations between the two measures. Nevertheless, these correlations were not as high as the ones found in test development and study 1 (Chapter 5 and Chapter 6). This may be because study 2 was conducted worldwide and participants might have had different opinions about what each level meant on the Likert Scale (i.e., what a 'beginner' level meant). Overall, the correlation between C-test scores and self-perceived proficiency is of only secondary interest for the test purpose in this chapter. If it is a main interest, future researchers should use a more specific and clear self-assessment questionnaire consisting of "can-do" statements adapted from language frameworks such as CEFR.

On the other hand, the correlations between C-test and TYS are of fundamental interest in this chapter given that the TYS is taken as the gold standard measurement of test takers' proficiency. Assumptions 11 and 12 were accepted based on the correlations between C-test scores and TYS level, TYS total score, and TYS receptive (listening, reading) skills which were all towards the higher end of the correlations stated in Eckes and Grotjahn (2006). It is interesting that the Turkish C-test had the highest correlation with TYS listening skill given that C-test does not have an oral component, which was considered to be an issue for the sufficiency of the C-test to measure general language proficiency by participating teachers and examinees (see section 7.7.5.1) as well as some scholars (Chapelle & Abraham, 1990). On the other hand, the correlations between C-test scores and TYS productive (writing, reading) skills were towards the lower end of the correlational studies in Eckes and Grotjahn (2006). Despite this, the stronger correlation between C-test scores and TYS total score supports that the Turkish C-test relates to the general language proficiency in Turkish. As is known, this chapter investigates whether the Turkish C-test can predict overall performance in TYS rather than performance in the individual skill sections.

Assumption 13 about the predictive power of the Turkish C-test for TYS levels and total scores was also accepted looking at several regression analyses. The Turkish C-test was able to distinguish between all TYS levels except for C2 level. The C-test was found to be not difficult enough to distinguish C2 level from C1 level learners aligning with other studies showing the insufficiency of the C-test to discriminate among high-level learners (i.e., Grotjahn, 1987; Klein-Braley, 1985; Son, 2018). According to the regression with quadratic and cubic models, even though learners achieved the highest possible score on the C-test, they would not be able to

achieve C2 level on TYS. It would be possible with linear regression or ordinal regression model. However, linear and ordinal regressions would not be as effective as quadratic or cubic models in separating between below B2 and B2 levels probably due to their less flexible structure, which would be more problematic considering the main aim of using the Turkish C-test as a screening test was to see whether candidates were ready to take TYS. Therefore, cut scores on the C-test was set by using quadratic model and they were as following : Below B2 level between 0 and 68, B2 level between 69 and 94, C1 level equal to and above 95.

The accuracy rate of placing TYS candidates in the right levels based on the above stated cut scores was 68%. This is very similar to the findings of accuracy rates, ranging between 72% and 79%, in Papageorgiou and Cho (2014) when they determined the cut scores on TOEFL Junior Standard for placing students into correct ESL classes by using ordinal logistic regression. As in Papageorgiou and Cho (2014), the rate of false negative classifications was higher than the rate of false positives among misplaced test takers, which means that if misplaced, there is a higher chance of being placed below the actual level due to the test not being able to predict C2 levels. However, there were not any test takers who were placed in below B2 (unsuccessful at TYS) although they were successful in TYS showing that the Turkish C-test would not deter able candidates from taking the TYS. Qualitative findings also supported this finding showing that the C-test had motivating and encouraging impact on candidates. Overall, these cut scores should be used with caution given the small sample size and the low number of Below B2 level candidates in this study. A larger number of participants would be required for determining more accurate cut scores. Nevertheless, these cut scores provide a good start for predicting TYS levels. Therefore, the assumption 14 about the accuracy and sufficiency of cut scores to

predict TYS levels was ‘partially’ accepted since the cut scores couldn’t predict C2 level. One way to increase the predictive power of the C-test might be to get the same students to take multiple C-tests over a short period of time and then use the aggregated information to predict TYS.

Given that screening tests are used to identify learners at risk such as low-level learners, it was not more of a problem that C-test was not able to distinguish highest level learners. C2 level learners would not necessarily need a screening test. Furthermore, as revealed on the thematic analysis of qualitative findings, high level learners would prefer to practice language with more authentic ways such as talking to L1 speakers and reading newspaper. However, if some institutions required C2 level (i.e., a foreign ministry looking for a translator of Turkish) and there was a need for a test distinguishing among higher levels, the easier texts (i.e., T1, T3) in the Turkish C-test could be replaced with more difficult texts such as T12. Even if this was done, it is not known whether C-tests are in general suitable for higher level learners looking at the ceiling effects found in several other studies (i.e., Grotjahn, 1987; Klein-Braley, 1985; Son, 2018) (see section 8.3 for a more general discussion of this limitation). Therefore, it might prove more fruitful to investigate other test formats for higher level learners.

7.8.5 Decision

The evidence regarding the decision inference is based on the analysis of interviews and survey responses as well as all the previous statistical analyses conducted for each inference. Turkish C-test is yet to be as a screening test by TYS candidates in their test preparation to be able to fully explore its consequences.

Assumption 15 about the usefulness of the Turkish C-test to predict TYS levels and practice Turkish was ‘partially’ met. Teachers and candidates were

sceptical about the sufficiency of the Turkish C-test to predict TYS levels since the formats of the two tests were different, and in particular, C-test did not have an oral section. Nevertheless, as seen in the quantitative findings, the Turkish C-test had the highest correlation with the TYS listening section and was able to classify candidates under correct TYS levels with 68% accuracy. This scepticism was also found in study 1 while investigating researchers' perception of the C-test. As previously discussed in section 6.8.5, stakeholders' unfamiliarity with the C-test uses and construct is one of the biggest weaknesses of C-tests (Roever, 2018). Despite this general finding, TYS candidates were positive about using the Turkish C-test to practice their Turkish before TYS stating they found the test motivating and practical. They mostly thought that the C-test would be helpful to practice grammar, vocabulary, suffixes, and reading given it is quick and easy to have access. They also stated that the C-test was fitting for the structure of the Turkish language considering the agglutinative structure of Turkish and "the correct use of suffixes being key". Note that as a limitation, there were no interviewees below B2 level and only two of the instructors volunteered to participate in the interview.

Overall, the 8-text C-test reported in this chapter is useful for TYS candidates since it is freely available to TYS candidates in their test preparation, and candidates will benefit from taking a test that was aligned with TYS levels. If the test was used for higher-stakes decisions (i.e., placement of candidates in language classrooms) rather than as a method of self-evaluation, the study should be replicated with a larger sample size for more exact cut scores.

CHAPTER 8: GENERAL DISCUSSION AND CONCLUSION

8.1 Introduction

The present research aimed to validate two different uses of a newly developed Turkish C-test: (1) controlling the general language proficiency of Turkish L2 learners in SLA studies; (2) predicting readiness for TYS as a screening test. This was conducted through three empirical studies that developed a new Turkish C-test (Chapter 5), validated its uses as an SLA research instrument (Chapter 6) and a screening test for TYS (Chapter 7) by using Kane's (2006) argument-based approach to validation.

The results showed that the Turkish C-test was able to reliably distinguish across 4 different ability levels of Turkish L2 learners in the US and UK in study 1 providing evidence that it can be used to control the language proficiency. Also, none of the texts in the Turkish C-test favoured one of these groups (UK and US) over another. While there was a more equal distribution and a wider range of proficiency levels in study 1, most participants (94%) were B2 or above B2 level in study 2. Furthermore, while the most common L1 (61%) was English in study 1, it was a Turkic language (57%) coming from the same language family as Turkish in study 2. This led this application of the Turkish C-test to only spread learners across three different levels in study 2. Given that the aim of the second study was to investigate the screening potential of the Turkish C-test for lower level learners, this wasn't considered an issue. Both studies, however, showed stakeholders' scepticism toward the Turkish C-test reflecting the low validity of C-tests in the literature (i.e., Legenhausen, 1989; McBeath, 1989; Sigott, 2004; Sumbling et al, 2014). Despite this, researchers and learners were positive in using the C-test in their research or TYS preparation.

8.2 Contributions

8.2.1 Contributions to research

The research reported in this thesis makes a number of important contributions to the literature of the argument-based approach to validation and C-tests. This dissertation included two different validation studies (chapters 6 and 7) for two different test uses using argument-based approach since each test use and claim required different evidences. Validation processes involved a series of assumptions and evaluation questions that were answered by collecting specific evidences. They exemplified how much of what kind of evidence are needed for the suggested test uses in an argument-based approach. In validation study 1 (Chapter 6), the Turkish C-test was claimed to be used to control for language proficiency in SLA studies when a heterogenous sample of L2 learners is recruited in terms of language proficiency and the language proficiency may have an effect on the independent or dependent variable as a control variable (i.e., controlling language proficiency while investigating the effect of L1 topic knowledge on L2 reading comprehension). Therefore, for instance, for the scoring inference of this test claim, evidence regarding the discriminative power of the C-test to distribute learners along a wide continuum of abilities and elicit a wide range of scores was sufficient. In validation study 2 (Chapter 7), the Turkish C-test was claimed to be used as a screening test to predict test taker levels on TYS. Therefore, to the contrary of the study 1 where the C-test was not compared with another gold standard of proficiency, it was essential to investigate the relationship of the C-test with TYS and set cut scores on the C-test based on the TYS levels for the extrapolation inference of this test claim. Thus, this dissertation showed how assumptions and evaluation questions are tailored for different test uses. It exemplified the implementation of an argument-based approach in a low-stake

context contrary to many studies using it for higher-stakes decisions such as university admissions.

Another contribution of this dissertation relates to the C-test literature. This dissertation is unique in showing how the challenges resulting from these morphological features of Turkish are addressed in developing a Turkish C-test for adult learners of Turkish in Chapter 5. Thus, this dissertation is the first study to specify the development stages of a Turkish C-test step by step with language specific factors. Furthermore, it is the first attempt to design a Turkish C-test which covers a wide range of proficiency levels and distinguishes between adult Turkish L2 learners of different proficiency levels. Overall, it shows that C-test method is applicable to Turkish language and contributes a new Turkish C-test validated as a research instrument and screening test to the existing literature of C-tests. It provides important guidance that should be taken into consideration in the design of C-tests in the Turkish language. Future researchers can benefit from these steps and guidance when developing their own C-tests for different populations (i.e., refugees, children). This dissertation also showed that 80% accuracy rate can be taken as criterion in agglutinative languages such as Turkish while piloting C-test texts with L1 speakers to come up with higher-level texts aiming for advanced level learners. Taking 90% accuracy rate produced a test that was not challenging enough for a group of advanced level learners as observed in Chapter 5. Including one new text using 80% accuracy rate was able to produce a test addressing advanced level learners as seen in study 1 in Chapter 6. However, including two news texts with this difficulty level, rather than one, could be better for advanced Azerbaijani learners of Turkish in study 2.

8.2.2 Contributions to practice

The developed Turkish C-test addresses the problems regarding the lack of validated measurement instruments in Turkish language that are freely available to researchers and learners. Several claims can be made about the uses of the Turkish C-test.

First it can be reliably used to control the language proficiency of adult Turkish L2 learners by SLA researchers in the US and UK when the language proficiency is a control variable (see Chapter 6). It can also fit within the time limitations of SLA researchers since it is quick and easy to administer and score. However, the test may not be used by SLA researchers when the language proficiency is a dependent variable in research studies since sufficient evidence was not collected to investigate the relation of the Turkish C-test with standardized measures of language proficiency in Study 1.

Another use of the Turkish C-test is that TYS candidates can take the test online anytime anywhere when they are preparing for the TYS and get an estimate of their TYS levels (except C2 levels) since cut scores on the Turkish C-test were determined based on TYS levels (below B2, B2, C1, C2) (see Chapter 7). In this way, candidates can see whether they are ready to take the TYS (i.e., having a minimum of B2 or C1 level for university entry) since they fail the TYS if they get below B2 level. If they are not at least at B2 level, they may need more practice and preparation for the TYS. This can help TYS candidates save time, money, and energy. Although TYS price does not seem to be a very big amount (i.e., 100 euros for candidates in Germany), candidates may not want to pay the exam fee several times if they don't get the level required for their job or university entry applications (see section 3.2.2 for details about the TYS). The Turkish C-test was found to be effective in distinguishing between all TYS levels except for the highest C2 level. Nevertheless,

given that screening tests are used to identify learners at risk such as low-level learners, it was not more of a problem that C-test was not able to distinguish highest level learners. As revealed on the thematic analysis of qualitative findings, Turkish C-test helped candidates to realise their weaknesses in Turkish, such as suffixes and reading academic topics although higher level candidates would prefer to practice language with more authentic ways such as talking to L1 speakers and reading newspaper (see section 7.7.5.1.2). Nevertheless, before publicly promoting the C-test as a screening test for TYS, a replication study conducted with a larger sample size would be necessary (see the discussion in the following section 8.3)

8.3 Limitations

Although this dissertation has made significant contributions to research and practice, it is important to discuss several methodological limitations. Regarding data collection and sampling, online data collection in validation studies allowed reaching a larger sample size compared to paper and pencil data collection in the initial investigation of the test development (see section 4.4 for details about data collection). However, there were several limitations that came with online test administration in validation studies.

First, the data was based on self-selection in that only volunteering people participated in the online study. While the sample was still very heterogenous in validation study 1 in terms of language proficiency levels (as initially determined by institutional status) and language background characteristics (i.e., age of learning Turkish, months of residence in Turkey), only 6% of the sample in validation study 2 involved test takers below B2 level as indicated by TYS results. Thus, for validation study 2, a more balanced sample size with a higher number of failed students would have been ideal. As Drackert (2016) also stated, beginner level learners might be

intimated by tests that cover the whole L2 proficiency range and have less motivation to participate. This issue could be addressed with a follow-up study using all TYS takers as the population and then using sampling weights to reweight any given sample of test takers back to the population. This work could be done in conjunction with the TYS. Though the exam institute might not want to for commercial reasons.

Another issue resulted from that participants could take the study anytime and anywhere as long as they had internet connection. Although this gave participants convenience, finalizing data collection took more time than expected since people did the test on their own time. Therefore, the study links were open to participants for a period of 8 months. Furthermore, there was less control on them to prevent consulting to language sources such as dictionaries. In order to minimize the possibility of cheating and ensure that learners do the test in one session, the test was administered with time limitations as in other unproctored internet testing (UIT) studies (i.e., Makransy & Glas, 2011; Nye, Do, Drasgow, & Fine, 2008). Furthermore, since the Turkish C-test was designed to be used for research and screening purposes rather than any high stakes decisions (i.e., placement, grades, recruitment) and, thus, did not have any direct consequences on test takers, it was expected that test takers did not have any incentive to cheat. On the other hand, learners may also have had less incentive to do their best on the test since the test did not have a direct consequence on them. Reimbursing participants for their time and efforts with gift cards might have motivated them to do their best.

Regarding the sample size, although the number of participants in both validation studies are small for a large-scale validation study, they are similar to other validation studies in LCTL which uses inferential statistics and Rasch analysis (i.e., Drackert, 2016; Son, 2018). It is acknowledged that Rasch analysis requires at least

10 observations per category for polytomous items (see section 4.5.2.1). This means that 210 participants would be required for both validation studies considering 0 to 20 polytomous rating scale of C-test texts. However, this size is very difficult to reach in LCTL such as Turkish. Thus, a replication study conducted over a larger period time and involving a larger sample size would be necessary before publishing the C-test online open to thousands of learners to provide them immediate information about their levels (Chapelle, Jamieson, & Hegelheimer, 2003). This dissertation provided the preliminary findings and showed that the C-test would be useful to provide estimates of the CEFR-aligned TYS levels.

Regarding the instruments, the Turkish C-test was not able to distinguish between learners with higher ability levels (C1 and C2) in study 2. Therefore, if there is a precise need to discriminate among higher ability levels (i.e., ministries recruiting translators of Turkish), new texts of higher difficulty such as Text 12 might be included or replace the easier texts in the test. However, as discussed in section 2.3.2.2, it is unclear whether even the most difficult C-tests can distinguish well between high-level proficient learners. It is worth noting that this ceiling effect was commonly observed in other studies investigating the discriminative power of C-tests as well as cloze tests (i.e., Grotjahn, 1987; Klein-Braley, 1985; Oller & Conrad, 1971; Tremblay, 2011; Son, 2018). As explained in section 2.3, this common finding might be attributed to that C-tests and cloze tests estimate proficiency in written modality and might not closely reflect L2 learners' oral/aural proficiency. However, the findings of the study 2 contradicted this interpretation in that the C-test had the highest correlations with TYS listening. Thus, depending on the intended uses, alternative forms of testing might be investigated for higher level learners.

Regarding the adapted validation framework, argument-based approach provided enough flexibility and guidance to form the interpretive arguments for the suggested test uses and. However, the missing point in the argument-based approach was the unclarity about who is supposed to judge the clarity, sufficiency and relevance of the interpretive argument. For example, if the argument-based approach is used by SLA researchers or language institutions, they may not have the required assessment literacy and methodological expertise to employ the argument-based approach.

8.4 Suggestions for Future Research

The research reported in this thesis offers a range of opportunities for future research. Some suggestions have already been made in response to the limitations stated above, which included a follow-up study using sampling weights, a replication study with a larger sample size, and a new test where there are one or two more difficult texts replacing the easier ones for higher-level learners. In addition to these suggestions, familiarising test takers with the C-test format and using gap-level factors to determine the difficulty of C-test texts might be considered.

First, it is worth noting again that almost all participating learners of this dissertation were unfamiliar with the C-test structure and it was the first time they took such a test in this format. They were only informed about the test through written instructions and an example sentence since there was no invigilator or test administrator in test locations. Thus, test taker unfamiliarity with the test format and test content was one of themes revealed in the thematic analysis of interviews and surveys. In the future, it would be useful to familiarize learners with C-test format through different classroom tasks or more online examples before they take the test. One may argue that familiarity with the test format may also weaken the relationship

between the test score and the measured construct through practice test taking. However, test takers are already familiarised with the structure of TYS before they take it. In a similar vein, being familiar with the C-test might help to avoid the initial 'bewilderment' test takers may experience when they see the test format.

Another suggestion would be developing an adaptive version of the Turkish C-test. The differing range of difficulty across texts in one C-test and its potential to discourage learners came up during researcher interviews. A computer adapted version of the C-test might be used in order to match the test level to the test taker level and prevent the potential test taker fatigue as well as discouragement during the test so that test takers received the texts adapted for their levels earlier on.

Regarding the test itself, paragraph difficulty was the main criterion to assess the difficulty of the Turkish C-test because data analysis was conducted at the text level considering C-test texts as superitems. Paragraph level factors (beyond word and sentence level) were also found to explain 92% of variance in text difficulty while gap-level factors (word and sentence level) explained only 8% of the variance (Khoshdel et al, 2016). Nevertheless, in future research, it would be useful to examine Turkish C-test difficulty from both micro-level (i.e., word familiarity, cognateness, phonetic complexity) and macro-level perspective (i.e., inter-gap dependency, type-token ratio) suggested by Beinborn et al. (2014). By doing so, it might be clearer what kinds of texts and items can be used for complete beginner level and very advanced level Turkish language learners since it has been most challenging to come up with texts for these levels of learners.

As discussed in section 5.7, future research should also investigate the importance of some subjective decisions taken based on reasonable justifications. Researchers can compare C-tests with the 20 gaps vs. 25 gaps, ordering of C-test texts

in a mixed way or following an increasing level of difficulty, power C-tests vs. speeded C-tests.

8.5 Final Remarks

This dissertation developed and validated a Turkish C-test as a short-cut measure of general language proficiency that can be used to control the language proficiency of adult L2 learners of Turkish in SLA research and to predict candidates' performance and exam readiness for TYS. Kane's argument-based approach to validation was used to develop an interpretive argument for the suggested test uses and evaluate the strength and accuracy of the assumptions. The assumptions were found to be adequate to support the suggested test uses.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Alderson, J. C. (1978). *A study of the cloze procedure with native and non-native speakers of English* (Doctoral Dissertation). University of Edinburgh: Edinburgh, UK.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13(2), 219-227.

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Andrich, D. (1978) Application of psychometric rating model to ordered categories which are scored with successive integers *Applied Psychological Measurement*, *2*, 581–594.

Arras, U., Eckes, T., & Grotjahn, R. (2002). C-Tests im Rahmen des 'Test Deutsch als Fremdsprache' (TestDaF): Erste Forschungsergebnisse/ C-tests within the 'Test of German as a foreign language' (TestDaF): preliminary research findings. In R. Grotjahn, R. (Ed), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen/ The C-test: theoretical foundations and practical applications* (p. 175-209). Frankfurt/M.: Lang

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System*, 29 (2), 209-219.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, England: Oxford University Press.

Baghaei, P. (2008a). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from Current Research* (pp. 101-112). Frankfurt/M.: Lang

Baghaei, P. (2008b). The effects of the rhetorical organization of texts on the C-Test construct: A Rasch modelling study. *Melbourne Papers in Language and Testing*, 13(2), 32-52.

Baghaei, P. (2010). A comparison of three polychotomous Rasch models for super-item analysis. *Psychological Test and Assessment Modeling*, 52, 313-322.

Baghaei, P., & Grotjahn, R. (2014). Estimating the construct validity of conversational C-Tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling*, 56(1), 60-82.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R.(2007). *The English Lexicon Project. Behavior Research Methods, 39,* 445-459

Baur, R. S., & Meder, G. (1994). C-tests zur Ermittlung der globalen Sprachfähigkeit im Deutschen und in der Muttersprache bei ausländern Schülern in der Bundesrepublik Deutschland. In R. Grotjahn (Ed), *Der C-test: Theoretische Grundlagen und Praktische Anwendungen* (pp. 151-178). Bochum: Brockmeyer.

Bayyurt, Y., & Martı, L. (2016). The use of suggestion formulas in L2 Turkish. In A. Gürel (Ed.), *Second language acquisition of Turkish* (pp 195-219). Amsterdam: John Benjamins Publishing.

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2, 517–529.

Boonsathorn, S. (1987). *C-Tests, proficiency, and reading strategies in ESL* (Unpublished PhD dissertation). University of Alberta, Canada.

Brennan, R. (2001). *Generalizability theory*. New York: Springer-Verlag.

Braun, V. & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3 (2), 77-101.

Brown, J. D. (1989). Cloze item difficulty. *Journal of the Japan Association of Language Teachers*, 11(1), 46–67.

Brown, J. D. (2008). Testing context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly*, 5(4), 275 -312

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to

second language teaching and testing. *Applied Linguistics*, 1, 1-47.

Caprez-Krompák, E., & Gönç, M. (2006). Der C-test im Albanischen und Türkischen Theoretische Überlegungen und empirische Befunde. In R. Grotjahn (Eds.), *Der C-Test: Theorie, Empirie, Anwendungen* (pp. 243-260). Frankfurt/M: Lang.

Carroll, J. B., (1961). Fundamental considerations in testing for English language proficiency of students. In H. B. Allen (Ed.), *Teaching English as a Second Language* (pp. 364-372*)*. New York, NY: McGraw Hill.

Carroll, J. B., (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14, 3-22.

Chapelle, C. A. (1994) Are C-tests valid measures for L2 vocabulary research? *Second Language Research,* 10(2), 157–187.

Chapelle, C.A., & Abraham, R.G. (1990). Cloze method: what difference does it make? *Language Testing,* 7, 121–46.

Chapelle, C. A., Enright, M., & Jamieson, J., (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.

Chapelle, C. A., Jamieson, J. , & Hegelheimer, V. (2003). Validation of a Web-based EFL test. *Language Testing*, 4, 409-439.

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol

3, pp. 1079–1097). John Wiley and Sons, Inc.

Chihara, T., Cline, W. D., & Sakurai, T. (1996). If the cloze test is a question, is the C-test the answer? In R. Grotjahn (Ed.) *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 183-195). Bochum: Brockmeyer.

Chiu, C. W. C. (2001). *Scoring performance assessments based on judgments: Generalizability theory*. New York: Kluwer.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice,* 26(2), 212-225.

Cizek, G., Rosenberg, S., & Koons, H., (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement,* 68 (3), 397-412.

Cleary, C. (1988). The C-Test in English: Left-hand deletions. *RELC Journal*, 19, 26–35.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York: Academic Press.

Cohen, A.D., Segal, M., & Weiss Bar-Siman-Tov, R. (1985). The C-Test in Hebrew. In C. Klein-Braley, & U. Raatz (Eds.), *C-Tests in der Praxis,* (pp. 121-27). Bochum: AKS-Verlag.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press & Council of Europe.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd. ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches.* (2nd ed.). Thousand Oaks, CA: Sage

Crewell, J. W., & Zhou, Y. (2016). What is mixed methods Research? In A. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second Language Assessment and Mixed Methods Research.* Studies in Language Testing, 43 (pp. 35-50). Cambridge, UK: Cambridge University Press.

Crotty, M. (1998). *The foundations of social research: meaning and perspective in the research process,* London: Sage.

Daller, H., Treffers-Daller, J., Ünaldi, A., & Yildiz, C. (2002). The development of a Turkish C-test. In J. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University Language Testing and the C-test* (pp. 187-199). Bochum: AKS Verlag.

Daller, H., & Phelan, D. **(**2006). The C-test and TOEIC® as measures of students' progress in intensive short courses in EFL. In R. Grotjahn, (Ed), *Der C-Test: Theorie, Empirie, Anwendungen/ The C-test: theory, empirical research, applications* (pp. 101-119). Frankfurt am Main: Peter Lang.

Dirgin, J. (2014). *An Overview of the ILR Skill Level Descriptors for Proficiency Test Development and Evaluation*. Paper presented at the annual meeting of the Language Flagship in California, USA.

Do, B. (2009). Research on unproctored internet testing. *Industrial and Organizational Psychology*, 2, 49–51.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-Test amongst Hungarian EFL

learners. *Language Testing*, 9, 187-206.

Drackert, A. (2016). *Validating Language Proficiency Assessments in Second Language Acquisition Research*. Frankfurt: Peter Lang.

Dunlea, J. (2015). *Validating a set of Japanese EFL proficiency tests: Demonstrating locally designed tests meet international standards* (Doctoral Dissertation). University of Bedfordshire: Bedfordshire, UK.

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung [Rasch models for C-tests]. In R. Grotjahn (Ed.) *Der C-Test: Theorie, Empirie, Anwendungen* [The C-test: theory, empirical research, applications] (pp. 1-44). Frankfurt am Main: Peter Lang.

Eckes, T. (2007). Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen. *Diagnostica*, 53(2), 68–82.

Eckes, T (2011) Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53*, 414–439.

Eckes, T. (2014). Die onDaF-TestDaF-Vergleichsstudie: Wie gut sagen Ergebnisse im onDaF Erfolg oder Misserfolg beim TestDaF vorher? In R. Grotjahn (Ed.) *Der C-test: Aktuelle Tendenzen*. [The C-test: Current trends] (pp. 137–162). Frankfurt am Main: Peter Lang.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, 23 (3), 290-325.

Eckes , T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education*, 28(2), 85-98.

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based english language screening tool. *Language Assessment Quarterly*, 5(3), 173-194.

Ellis, D. P., & Ross, S. J. (2014). Item Response Theory in Language Testing. In A. J. Kunnan (Ed.) *The Companion to Language Assessment* (1262-1281).West Sussex, UK: John Wiley & Sons, Inc.

Ergül, E. (2017, June). Enrollment Survey. *American Association of Teachers of Turkic Languages*, 6-10. Retrieved from https://www.international.ucla.edu/media/files/AATT-June-2017--tk-4jc.pdf

Feldmann, U., & Stemmer, B. (1987). Thin_ aloud a_ retrospective da_ in c-te_ taking: differ_ languages – diff_ learners – sa_ approaches? In C. Faerch, & C. Kasper (Eds.), *Introspection in second language research* (pp. 251-267). Clevedon: Multilingual Matters.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), (pp. 105–146). Washington, DC: The American Council on Education/Macmillan.

Field, A. (2013). *Discovering statistics using SPSS*, (4th ed.). Los Angeles, CA: Sage.

Fulcher, G. (2014). Philosophy and language testing. In A. J. Kunnan (Ed)., *The Companion to Language Assessment* (pp. 1431-1451). London: Wiley-Blackwell.

Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.

Gaillard, S. (2014). *The Elicited imitation task as a method for French proficiency assessment in institutional and research settings* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

Green, A. (2012). Placement testing. In C. Coombe, B. O' Sullivan, P. Davidson, & S. Stoynoff (Eds.), *The Cambridge guide to language assessment* (pp. 164–170). Cambridge: Cambridge University Press.

Grotjahn, R. (1987). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In R. Grotjahn, C. Klein-Braley, and D. K. Stevenson (Eds.), *Taking their measure: The validity and validation of language tests* (219-253). Bochum: Brockmeyer.

Grotjahn, R. (2002). 'Scrambled' C-tests: Eine Folgeuntersuchung. In R, Grotjahn (Ed), *Der C-test*: *Theoretische Grundlagen und praktische Anwendungen* (pp 211-225). Bochum: AHS-Verlag

Grotjahn, Rüdiger. (2017). The electronic C-test bibliography: version 2017 (last update: April 10, 2017). Available at: http://www.C-test.de.

Grotjahn, R., & Allner, B. (1996). Der C-Test in der Sprachlichen Aufnahmeprüfung an Studienkollegs für ausländische Studierende an Universitäten in NordrheinWestfalen. In Rüdiger Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 3, pp. 279–335). Bochum: Brockmeyer.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, 1, 1-10.

Gürel, A. (2016). *Second language acquisition of Turkish*. Amsterdam: John Benjamins Publishing.

Haertel, E. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed.), (pp. 65–110), Westport, CT: American Council on Education and Praeger.

Hankamer, J. (1989). Morphological parsing and the lexicon. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 392–408). Cambridge, MA: MIT Press.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, 11(2), 152-169.

Harsch, C. & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, 33(4), 555-575.

Hastings, A. (2002). In defense of C-testing. In R. Grotjahn (Ed), *Der C-test*: *Theoretische Grundlagen und praktische Anwendungen* (pp 11-29). Bochum: AHS-Verlag.

Henning, G. (1987) *A guide to language testing: development, evaluation, research*. Newbury House, Cambridge MA.

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S., & Schedl, M. (2008). Prototying a new test. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds), *Building a validity argument for the Test of English as a Foreign Language* (pp. 187- 225). New York, NY: Routledge

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal, 91,* 663–667.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*, 229–249.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, 15, 422-433.

Hulstijn, J. H. (2015). Language proficiency in native and non-native speakers: Theory and research. Amsterdam: John Benjamins Publishing Company.

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.) *Sociolinguistics* (pp. 269-293). Harmondsworth, UK: Penguin Books.

Huhta, A. (1996). Validating an EFL C-test for students of English philology. In R. Grotjahn (Ed), Der C-Test: Theoretische Grundlagen und praktische Anwendungen / The C-test: theoretical foundations and practical applications (pp. 197-234). Bochum: Brockmeyer.

Interagency Language Roundtable. (1985). Interagency Language Roundtable skill level descriptions – Reading. Retrieved October 15, 2015, http://www.govtilr.org/skills/ILRscale4.htm.

Jafarpur, A. (2002). A comparative study of a C-Test and a cloze test. In R. Grotjahn (Ed), Der C-Test: Theoretische Grundlagen und praktische Anwendungen / The C-test: theoretical foundations and practical applications (pp. 31-51). Bochum: AKS-Verlag.

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 243-274). New York: Routledge.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly*, 24 (1), 61-83.

Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing*, 8 (1), 1-22.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14-26.

Kane, M. T. (1990). An argument-based approach to validation. ACT Research Report Series 90-13. Iowa City, IA: American College Testing.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement:Issues and Practice*, 21 (1), 31–41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, 2, 135–170.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.), (pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, 29(1) 3-17.

Kane, M. T. (2013). Validating the interpretations and uses of Test Scores. *Journal of Educational Measurement*, 50, 1-73.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy, & Practice*, 23(2), 198-211.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18, 5-17.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading.* Studies in Language Testing, 29. Cambridge: Cambridge University Press.

Khoshdel, F., Baghaei, P., & Bemani, M. (2016). Investigating factors of difficulty in C-tests: A construct identification approach. *International Journal of Language Testing,* 6 (2), 113-122.

Klein-Braley, C. (1981). *Empirical investigations of cloze tests* (Doctoral Dissertation). University of Duisburg: Duisburg, Germany.

Klein-Braley, C. (1984). Advance prediction of difficulty with C-Tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing* (Vol 7, pp. 97-112. Colchester: University of Essex.

Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the construct validation of authentic texts. *Language Testing*, 1 (1), 76-104.

Klein-Braley, C. (1994). *Language testing with the C-Test: A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty* (Unpublished Higher Thesis). Department of Linguistics and Literature, University of Duisburg: Germany.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, 14 (1), 47-84.

Klein-Braley, C., & Raatz, U. (1982). Der C-Test: Ein neuer Ansatz zur Messung von allgemeiner Sprachbeherrshung. *AKS-Rundbrief*, 4, 23-37.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing,* 1 (2), 134-146.

Knoch U., Elder C., O'Hagan S. (2016). Examining the Validity of a Post-Entry Screening Tool Embedded in a Specific Policy Context. In Read J. (Ed) *Post-admission Language Assessment of University Students,* (pp. 23-42). Springer, Cham, Switzerland: Springer

Lado, R. (1961). *Language Testing*. New York, NY: McGraw-Hill.

LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, 19(4), 673-687.

Legenhausen, L. (1989). Zur face validity der C-tests: Lehrer und Schülerurtelie. In Finkenstaedt, T., & Schröder, K. (Eds). *Zwischen Empirie und Machbarkeit: Erstes Symposium zum Bundeswettbewerb Fremd-sprachen* (pp. 70-81). Augsburg: Universität.

Lewis, G. (2000) *Turkish Grammar* (2nd edition). Oxford: Oxford University Press

Lee-Ellis, S. (2009). The development and validation of a Korean C-test using Rasch Analysis. *Language Testing*, *26*, 245–274.

Li, Z. (2015). *An argument-based validation study of the English Placement Test (EPT) – Focusing on the inferences of extrapolation and ramification* (Doctoral dissertation). Iowa Stata University, Ames, IA.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre J. M. (2002) Understanding Rasch measurement: Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3(1), 85-106.

Linacre, J. M. (2017a) Sample size and item calibration [or person measure] stability. Retrieved May 25, 2019 from https://wwwraschorg/rmt/rmt74mhtm

Linacre, J. M. (2017b). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com

Linacre, J. M. (2018). *A user's guide to FACETS*. Chicago, IL: Winsteps.com

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Makransky, G., & Glas, C.A.W. (2011). Unproctored internet test verification: Using adaptive confirmation testing. *Organizational Research Methods*, 14 (4), 608-630.

Masters, G. N. (1982) A Rasch model for partial credit scoring *Psychometrika*, *47*(2), 149–174.

McKay, T., & Abedin, N. (2018). Developing a C-test for Bangla. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 61-89). Frankfurt am Main, Germany: Peter Lang.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McBeath, N. (1989). *C-Tests in English: Pushed beyond the original concept? RELC Journal*, *20*(2), 36–41.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed, pp. 13-103). New York: American Council on Education and Macmillan.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series,* 2, 1-18.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.

Montrul, S. (1997). Transitivity alterations in second language acquisition research: A crosslinguistic study of English, Spanish, and Turkish (PhD Dissertation). McGill University.

Montrul, S. (2016). The causative/inchoative morphology in L2 Turkish under the feature reassembly approach. In A. Gürel (Ed.), *Second language acquisition of Turkish* (pp 107-133). Amsterdam: John Benjamins Publishing.

Negishi, M. (1987). The C-test: an integrative measure? *IRLT Bulletin* 1, 3–26.

Newton, P. E., & Baird, J. A. (Eds.) (2016). Validity [Special issue]. *Assessment in Education: Principles, Policy, & Practice*, 23 (2).

Norris, J. M., (2006). Development and evaluation of a curriculum-based German C-test for placement purposes. In R. Grotjahn (Ed.), *The C-test: Theory, Empirical Research, Applications* (pp. 45-83). Frankfurt: Peter Lang.

Norris, J. M. (2008). *Validity evaluation in language assessment*. Frankfurt, Germany: Peter Lang.

Norris, J. M. (2018). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research*. Frankfurt am Main, Germany: Peter Lang.

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass, & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). New York: Routledge.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. London, UK: Longman.

Oller, J. W. Jr., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning*, 21, 183-195.

O'Sullivan, B., & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: theories and practices* (pp. 13-32). Oxford: Palgrave

Macmillan.

Özçelik, Ö. (2011). *Representation and acquisition of stress: The case of Turkish.* (Doctoral Dissertation). McGill University.

Özçelik, Ö., & Sprouse, R. A. (2016). Vowel harmony in English-Turkish interlanguage. In A. Gürel (Ed.), *Second language acquisition of Turkish* (pp 49-72). Amsterdam: John Benjamins Publishing.

Öztopçu, K. (2009). *Elementary Turkish: A complete course for beginners*. Ankara: Kebikeç Yayınları-Sanat Kitabevi.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL Junior Standard scores for ESL Placement decisions in secondary education, *Language Testing*, 31(2), 223-239.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: Test usefulness evaluation, *Language Assessment Quarterly*, 7 (2), 137-159.

Popper, K. R. (1963). Conjecture and refutation: The growth of scientific knowledge. New York: Basic Books.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, Vol. 7. Language testing and assessment* (2nd ed.), (pp. 53–68). Dordrecht, The Netherlands: Kluwer.

Purpura, J., Brown, J., & Schoonen, R. (2015). Improving the validity of quantitative measures in second language research. *Language Learning*, 65 (Supplement 1), 36-73.

Raatz, U. (1985). Better theory for better tests? *Language Testing*, 2, 60-75.

Raatz, U., & Klein-Braley, C. (1982). The C-test – a modification of the cloze procedure. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing IV* (pp. 113-138). Colchester: University of Essex, Dept. of Language and Linguistics.

Raatz, U. & Klein-Braley, C. (1985). How to develop a C-Test. *Fremdsprachen und Hochschule*, 13/14, p. 20-22.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education*, 1-14.

Roover, C. (2008). Developing C-tests across eight languages: Discussion. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 295-304). Frankfurt am Main, Germany: Peter Lang.

Saito, Y. (2003). The use of self-assessment in second language assessment. *TESOL Web Journal.*

Sağın Şimşek, Ç. (2006). *Third language acquisition: Turkish-German bilingual students' acquisition of English word order in a German educational setting.* Münster: Waxmann.

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26(1), 5–30.

Schmidgall, J. E., Getman, E. P., & Zu, J. (2017). Screener tests need validation too: Weighing an argument for test use against practical concerns. *Language Testing*, Online first, https://doi.org/10.1177%2F0265532217718600.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.),

Review of Research in Education, (Vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(5–8), 13.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice*, 23(2), 268-280.

Sigott, G. (2004). *Towards identifying the C-test construct*. Frankfurt am Main: Peter Lang.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy and Practice,* 23 (2), p. 226-235.

Son, Y-A. (2018). *Measuring Heritage Language Learners' Proficiency for Research Purposes: An Argument-Based Validity Study of the Korean C-test*. (Doctoral dissertation). Georgetown University, Washington, DC.

Spolsky, B. (1969, September). Reduced redundancy as a language testing tool. Paper prepared to be read at the Language Testing section of the second International Congress of Applied Linguistics, Cambridge, England.

Spolsky, B. (1973). What does it mean to know a language; or how do you get somebody to perform his competence? In J. Oller, & J. Richards (Eds.), *Focus on the learner* (pp. 164-176). Rowley, MA: Newbury House.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (TOEFL Monograph Series No. MS-32). Princeton, NJ: ETS.

Strong-Krause, D. (2000). Exploring the effectiveness of self-assessment strategies in ESL placement. In G. Ekbatani, & H. Pierson (Eds.), *Learner-directed*

assessment in ESL (pp. 49- 73). Mahwah, New Jersey and London: Lawrence Erlbaum Associates.

Sumbling, M., Viladrich, C., Doval, E., & Riere, L. (2014). C-test as an indicator of general language proficiency in the context of a CBT (SIMTEST). In R. Grotjahn (Ed.) *Der C-test: Aktuelle Tendenzen*. [The C-test: Current trends] (pp. 53–108). Frankfurt am Main: Peter Lang.

Tabachnick, B.G., & Fidell, L.S. (2013) *Using Multivariate Statistics*. Pearson, Boston.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44,* 307–336.

Tinsley, T., & Board, K. (2014). Languages for the future. *British Councils*. Retrieved from https://www.britishcouncil.org/sites/default/files/languages-for-the-future-report-v3.pdf

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition, 33*, 339–372.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, England: Palgrave Macmillan.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao& S. Sinharay (Eds.), *Psychometrics* (pp. 45–79). Amsterdam: Elsevier Science.

Appendix 1: Ethical Approval and Ethics Form

Your online ethics application for your research project "Validating Language Tests in Second Language Acquisition Research and Educational Programs Through an Argument-Based Approach: The C-test in Turkish" has been granted ethical approval. Please ensure that any additional required approvals are in place before you undertake data collection, for example NHS R&D Trust approval, Research Governance Registration or Site Approval.

For your reference, details of your online ethics application can be found online here:

http://www.bristol.ac.uk/red/ethics-online-tool/applications/68942

Ethical issues discussed and decisions taken (see list of prompts overleaf):

Duygu Çavdar and I have met on June 3, 2018 and we discussed the 14 ethical issues suggested. The taken decisions are summarised below:

1. Researcher access/exit

First, we talked about the summary of the proposed research. Then, we discussed how to access participants. Participants will be invited to the study through e-mail invitations. If necessary, follow-up phone calls will be made.

2. Information given to participants

Information about study will be provided in the participant information sheet. The given information will be very concise not to affect student performance during the test.

3. Participants' right of withdrawal

It will be emphasized participation is entirely voluntary and withdrawal any time during the test is possible by quitting the test; however, participants will be encouraged to complete the test through instructions and the reward of £5 Starbucks gift card upon completion of the test. Furthermore, they will have an opportunity to enter £50 Prize Draw (Amazon Gift Card) upon their willingness to participate in the follow-up interviews.

4. Informed consent

Consent form at the end of the information sheet will also have information about the researcher, benefits of taking the test, absence of associated risks, participants' right to withdraw from the test any time. Information sheet will be given to participants before they take the test.

5. Complaints procedure

There will also be information about the procedure in the case of possible complaints. Participants will be able to contact the researcher, supervisor, and SoE ethics committee if they have any kind of complaints. Contact information of the researcher, supervisor, and SoE ethics committee will be provided.

6. Safety and well-being of participants/ researchers

We discussed how to protect the safety and well-being of participants as well as the researcher. There are no estimated risks associated with the study. The questionnaires and interviews do not involve any sensitive or personal topic. Furthermore, caution has been taken to avoid sensitive and controversial topics such as race or religion in the texts of the C-test. The texts were neutral in content.

7. Anonymity/confidentiality

To protect anonymity, participants will be assigned numbers and no data will be associated to any names. These numbers will be used in data collection and analysis. Any names or e-mail addresses will not be shared with third parties and used in data reporting.

8. Data collection

Data collection will be done online by using Qualtrics for surveys, in lab format and using audio recording via Skype or phone calls for interviews.

9. Data analysis

Data analysis will be conducted through software statistics tools which are SPSS and QSR NVivo 10. The audio recording will be transcribed by the researcher anonymously. Transcriptions will be checked to ensure they don't involve any personal and identifying information about participants. Participants will be given the opportunity to check the transcription of their own interview for correctness and clarity of meaning.

10. Data storage

Questionnaires and audio recordings and tests will be deleted after one year of data collection. However, SPSS and transcriptions will be stored in researcher's personal password-protected computer under 10 years to ensure researcher can have access to it for publication purposes.

11. Data protection act

Participants' questionnaires and tests will not be shared with third parties

12. Feedback

If participants want to learn study results, they will be able to reach the researcher through e-mail. They will be provided with a research summary after the study is completed if they request it.

13. Responsibilities to colleagues/ academic community

The study is expected to make a significant contribution to the research community. The test in question will be publicly available to researchers and language programs who want to use it.

14. Reporting of research

The write-up of the study will be in researcher's doctoral thesis. It will also be possibly presented in conferences and published in academic journals.

Signed: Merve Demiralp (Researcher)
Signed: Duygu Çavdar (Discussant)
Date: 03/06/2018

Appendix 2: Background Questionnaire for Test Development

Turkish Language Learning Background Questionnaire

Before the test, please fill out the following background information:

1. Age: _____

2. Gender: Male Female

3. Mother tongue (First language):
_____

4. Year in college:

Freshman Sophomore Junior Senior MA PhD Other_____

5. Major: _____

6. List the Turkish language classes you have taken and you are currently taking. Indicate if they are required and when & where you took/are taking them.

Course (number and title)	Required? (Yes or No)	When?	Where?

7. Please circle your proficiency level for Turkish in the following areas.

	Low Beginner 1	**2**	**3**	**4**	**Very Advanced 5**
Reading	1	2	3	4	5
Writing	1	2	3	4	5
Listening	1	2	3	4	5
Speaking	1	2	3	4	5
OVERALL Proficiency	1	2	3	4	5

8. At what age did you begin studying Turkish? _____

9. How long have you been studying Turkish? _____

10. Have you ever visited a Turkish-speaking country? _____ Yes _____ No

If yes, where and for how long?

Location	Length of visit

11. Outside of class, how many hours per week do you spend using Turkish?

_____

12. Do you have family members who speak Turkish? _____ Yes _____ No

If so, who (e.g., parents, grandparents,
etc.)?_____

13. Please list any other languages that you have previously studied:

Language	Length of study

14. Have you taken a Turkish proficiency test recently (e.g., an Oral Proficiency Interview or a placement test)? If so, what score did you receive (list test and score/rating)?

Thank you!

Now, please turn over and start the test.

Appendix 3: 11-text Turkish C-test for Test Development

Time at the beginning of C-test: _____

Instructions: Please read the instructions carefully before starting the test.

- The following 11 texts have been developed by deleting about second half of some words. Please fill in the blanks as shown in the example below.

  Example:
  Geçen Paz_____ önemli işle_____ bitirdim ve ta_____ için pl_____ yaptım.
  Geçen Pazar önemli işlerimi bitirdim ve tatil için plan yaptım.

- Pay attention to context, vocabulary, and grammar (e.g. subject-verb agreement, consistency of tense). Skim each text to get its meaning before filling out the blanks. In some texts, you need to read and understand all the text to complete the gaps.
- After you fill in the gaps in each text, read over the text and make sure your answers are consistent with the rest of the text such that you use appropriate verb tenses and personal pronoun markers.
- Some texts are considerably more difficult than others; but it is very important that you try your best to complete all blanks in all texts. Do not skip a text since you think it is difficult. However, don't worry if you are unable to fill out all of the blanks.
- The test is not timed; take the time you need for each text, but do not spend more than one hour on the full text.
- Do not use a dictionary or any other aids in completing the test.
- Be as accurate as possible (spelling counts), but do not be concerned if you do not know answer to any of the blanks.
- All the blank spaces are equal in length regardless of the length of the word.
- Please indicate the start time above and the end time on the last page.

Fill-in-the Blank Test

Text 1
Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir süpermar___ _____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

Text 2
Ben bir şirkette sekreterim. İşe he_____ sabah sa_____ dokuzda gidi_____. Şirketim evi_____ çok ya_____. Öğle yemeğ_____ şirkette yiy_____. Akşam iş_____ sonra alışve_____ yapıyorum. E_____ saat yedide otob_____ ile dönü_____. Biraz dinlen_____ ve ye_____ yiyorum. Yeme_____ sonra telev_____ seyrediyorum ya da ki_____ okuyorum. Ha_____ sonunda ba_____ spor yapı_____. Pazar günü genellikle sinemaya gidiyorum.

Text 3
Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis__ _____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

Text 4
İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san_____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslarar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____ __, sinema ve kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

Text 5
Bülent Ortaçgil, Türk gitarist, şarkıcı ve besteci. 1950 yıl_____ Ankara'da do_____. İlkokula Ankara'da baş_____ ve da_____ sonra İstanbul'a taş_____. Lise yılla_____ müzik çalışm_____ başladı. İl_____ gitarını b_____ akrabası al_____. Ortaçgil, Maarif Koleji'nde arkadaşl_____ beraber gi_____ çalmaya baş_____ ve çeş_____ müzik grup_____ kurdu. Far_____ farklı isim_____ amatör

mü_____ yapan b_____ grupların bir_____ de Damlalar ismini taşıyordu. Ortaçgil, grupta davul çalıyordu.

Text 6

Anakent Koleji, öğrencilerini geleceğe tam olarak hazırlamayı misyon edinen bir kurumdur. Okulumuzda yab_____ dil öğret_____ büyük ön_____ verilir. Öğren_____ İngilizce ve Alm_____ dillerini sı_____ ortamında ve aktiv_____ yoluyla öğren_____. Bu sür_____, hazırlık sınıfl_____ başlayarak 12. sın_____ kadar de_____ eder. D_____ öğrenme konus_____ en öne_____ etkinliğimiz Yab_____ Diller Kulü_____. Bu ku_____ küresel kon_____ hakkında uluslarar_____ çalışmalar yapar. Yabancı bir dili etkin bir şekilde kullanma adına kulüp çalışmalarımız önem taşır.

Text 7

Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında, motiva_____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

Text 8

Özellikle Batılı ülkeler, dillerini öğretmek için birçok çalışma yapmışlar ve hâlâ yapmaktadırlar. Türkiye'de de ben_____ çalışmalar gö_____ görülür ora_____ artmakta ve yaban_____ Türkçe öğr_____ çalışmaları yaygınla_____. Gerek yu_____ içinde, ger_____ yurt dış_____ pek ç_____ yerde Tür_____ öğretimi yapılm_____. Her ge_____ gün Tür_____ öğrenimine ol_____ talep artma_____. Bu doğru_____ nitelikli de_____ materyallerine ve progra_____ ihtiyaç duyulm_____. Bu materyaller, kurumların belirlediği programlar doğrultusunda oluşturulmaktadır.

Text 9

Koku alma duyusu tat duyusu ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

Text 10

Başarılı insanların en önemli özelliklerinden biri, harekete geçmektir. Adım at_____ ve hare_____ geçmek, si_____ hedeflerinize yakla_____. Ne var ki, ço_____ zaman karş_____ çıkan enge_____ sizi durd_____, çaresizliğe ve isteks_____ sürükler. B_____ hislerin ço_____ zaman işi_____ sevmenizle de ilgisi yok_____.

Motivasyonunuz dü_____ olduğu za_____, isteksizliğe ba_____ olarak üretken_____ de düşebilir. Tek_____ motive olun_____ kadar ge_____ zamanda ise, planlarınızın gerisinde kalabilirsiniz. Üretkenliğinizi ve motivasyonunuzu her zaman yüksek tutabilmek için zamanın bilincinde olmanız oldukça önemli.

Text 11

Türkiye'deki bilim kadınları hakkında yapılan hemen her araştırma, kimi ilginç olguların altını çizer. İlk ola_____, üniversitelerin far_____ kademelerinde y_____ alan kadın_____ oranı s_____ derece yüks_____. Sadece öğr_____ ya da asis_____ düzeyinde değ_____, öğretim üy_____ ve yöne_____ kadrolarındaki kadın_____ oranı da b_____ hayli kabar_____. Bunun yanısı_____, bilim kadı_____, kadın olma_____ dolayı he_____ hemen hiçb_____ ayrımcılığa uğramad_____ dile getirmektedirler. Bu saptamayı takip ettiğimizde 1930'lardan bu yana bir süreklilik buluruz.

Time at the end of C-test: _____

Thank you for your participation! Please turn over and complete the questionnaire.

Appendix 4: C-test Questionnaire for Test Development

Questionnaire About the Fill-in-the Blank Questions

1. Were the task instructions clear and concise? Was there anything that confused you?

2. Were the fill in the blank questions difficult to do? If so, what made them difficult?

3. Which text/texts were particularly difficult? Why?

4. Indicate the level of difficulty of each text, using the following scale. (Circle the appropriate number)

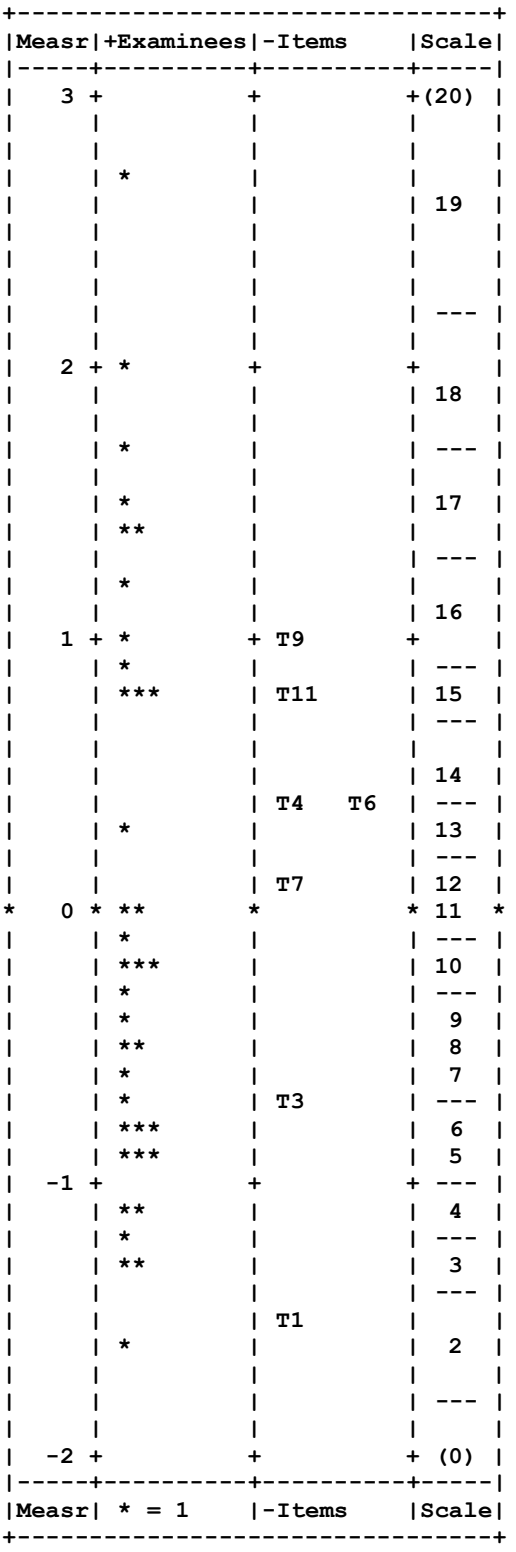
Very easy  Very difficult

Text 1	1	2	3	4	5
Text 2	1	2	3	4	5
Text 3	1	2	3	4	5
Text 4	1	2	3	4	5
Text 5	1	2	3	4	5
Text 6	1	2	3	4	5
Text 7	1	2	3	4	5
Text 8	1	2	3	4	5
Text 9	1	2	3	4	5
Text 10	1	2	3	4	5
Text 11	1	2	3	4	5

5. Among the 11 texts, have you seen any of them before? If so, where?

6. If you have any other comments, please feel free to provide them here.

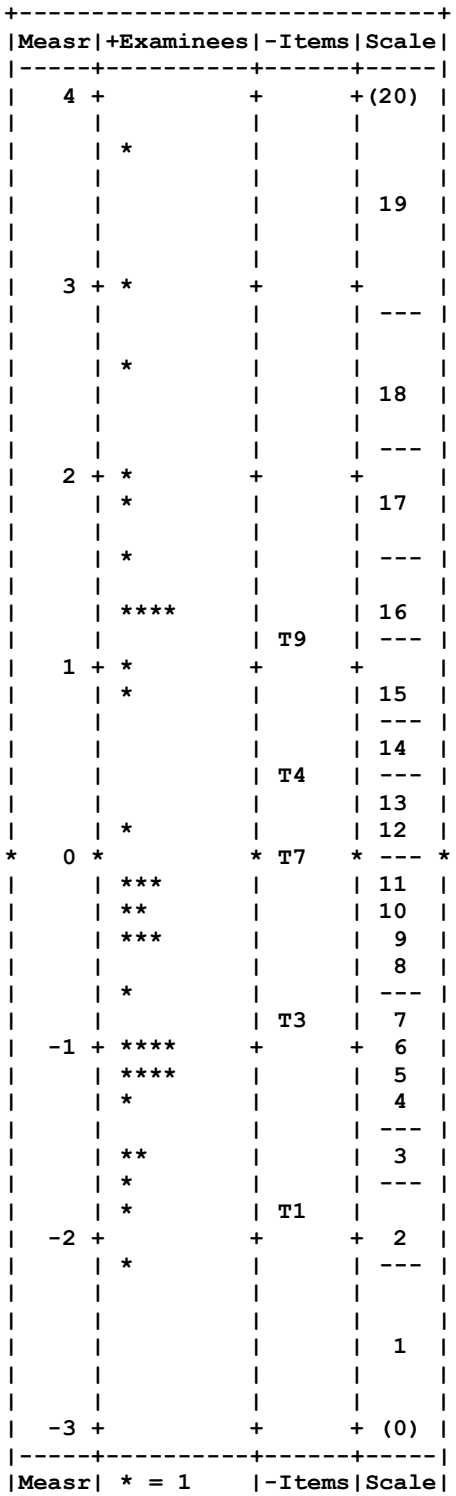
Appendix 5: Rasch Analysis with 7-text C-test in Test Development



Variance explained by Rasch model=90.86%
Separation=4.95; strata=6.86; reliability=.96

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.73	.87	1.31	1.14	.11	-1.48
T3	.83	1.02	1.04	.89	.08	-.68
T4	.96	1.59	.54	.55	.08	.36
T6	.90	1.05	.94	.86	.08	.40
T7	.90	1.19	.71	.74	.07	.05
T9	.91	1.27	.87	.80	.08	.95
T11	.91	.82	.86	.95	.08	.82

Appendix 6: Rasch analysis with 36 examinees in Test Development



Variance explained by Rasch model=92.67%
Separation=4.77; strata=6.69; reliability=.96

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.77	.29	1.25	1.08	.12	-1.83
T3	.84	1.01	1.07	.97	.09	-.88
T4	.93	1.29	.72	.72	.09	.41
T7	.90	.93	.90	.95	.09	.03
T9	.90	1.29	.72	.76	.10	1.14

Appendix 7: Background Questionnaire for L2 Learners in Validation Study 1

Turkish Language Learning Background Questionnaire

Before you take the test, please fill out the following background questionnaire.

Section 1: The following questions relate to your demographic information.

Q1 Please write your age.

_____

Q2 What is your gender?

- ○ Male
- ○ Female
- ○ Other
- ○ Prefer not to say

Q3 Which country are you currently working/studying in?

_____

Q4 Please write your mother tongue/first language (eg. English).

_____

Q5 Are you a heritage speaker of Turkish? (a person raised in a home where a non-majority language (eg. Turkish) is spoken is a **heritage speaker** of that language if she/he possesses some proficiency in it)

- ○ Yes
- ○ No

Q6 Are you a bilingual speaker of Turkish? **(bilingual speaker** means a person who has learned two or more languages relatively simultaneously during early childhood)

- ○ Yes
- ○ No

Q7 What is your highest completed degree of education?

- ○ High School
- ○ Undergraduate Degree
- ○ Master's Degree
- ○ PhD

Q8 What's your subject of study (or degree program/ discipline/ job)?

_____

Q9 Do you have any learning difficulties (eg. dyslexia) that can cause problems with reading, writing or spelling?

- ○ Yes
- ○ No

Q10 What is the type of your learning difficulty?

_____

Section 2: The following questions relate to your Turkish language learning experience.

Q11 Please list the names and the **levels** of the Turkish language classes that you have taken, and you are currently taking. Indicate if they are/were required and when & where you took or are taking them.

	Level	When?	Where?	Required
Course Name				
Course Name				
Course Name				
Course Name				
Course Name				

Q12 At what age did you begin studying Turkish?

_____

Q13 How long have you been studying Turkish? (in years and months)

_____

Q14 Have you ever visited a Turkish speaking country?

- ○ Yes
- ○ No

Q15 Please indicate where and how long you visited.

Location	Length of visit

Q16 Outside of class, how many hours per week do you spend using Turkish?

_____

Q17 Do you have family members who speak Turkish?

- ○ Yes
- ○ No

Q18 Please indicate the family members who speak Turkish (e.g., parents, grandparents, etc.)

_____

Section 3: The following question relates to your learning experience with other languages.

Q19 Please list any other languages that you have previously studied and the length of your study in years and months.

Language	Length of study

Section 4: The following questions relate to your Turkish language proficiency.

Q20 Have you taken any Turkish proficiency test recently? (e.g., an Oral Proficiency Interview, a placement test)

○ Yes

○ No

Q21 Please write the name of the Turkish test you have taken and your test score / rating.

_____

Q22 Please circle your self-perceived proficiency level for Turkish in the following areas.

	Beginner	**Elementary**	**Intermediate**	**Advanced**	**Very advanced**
Reading	○	○	○	○	○
Writing	○	○	○	○	○
Listening	○	○	○	○	○
Speaking	○	○	○	○	○
Overall Proficiency	○	○	○	○	○

Section 5: The remaining question relates to your experience with computers.

Q23 Please circle how comfortable you are with computer in the given conditions.

	Extremely comfortable	Moderately comfortable	Somewhat comfortable	A little bit comfortable	Not at all comfortable
using a computer	○	○	○	○	○
taking a test through a computer	○	○	○	○	○

Thank you for completing the questionnaire! *Now, please click on the next button to see the test instructions.*

Appendix 8: 6-text Turkish C-test for Validation Study 1

Instructions for the Test

Please read carefully. The test has not started yet.

In the following 6 texts, some letters are missing in a number of words. Please fill in the gaps by completing these words as shown in the example below.
Example:
Geçen Paz_____ önemli işle_____ bitirdim ve ta_____ için pl_____ yaptım.
Geçen Paz**ar** önemli işle**rimi** bitirdim ve ta**til** için pl**an** yaptım.

Don't panic! You may not be able to fill in all of the gaps; but it is very important that you try your best to complete all the blanks.

You can choose Turkish special characters (ç, ı, ğ, ö, ü, ş) **from the text box above the word that you are completing if necessary**. **Spelling counts,** so be as accurate as possible. Pay attention to context, vocabulary, and grammar.

You have a maximum of **30 minutes** to fill in all the gaps in all texts. If you haven't completed the test by the time limit, the test will be submitted automatically. So, it is recommended that you spend about 5 minutes per text. Your remaining time will be displayed on the screen.

You will be directed to the next text as soon as you have clicked on "submit" and "next". **You cannot go back once you click submit to a given text**.

After you fill in the gaps in each text, **read over the text to check there are no typos and make sure your answers are consistent with the rest of the text** such that you use appropriate verb tenses and personal pronoun markers. Your estimated score will be displayed at the end of the test. **Do not forget to complete the short participant feedback survey at the end.** Do **not use a dictionary or any other aids** in completing the test. Now, please tick the boxes below and click on "START THE TEST" button when you are ready.

☐ I confirm that I cannot go back to the texts that I have submitted

☐ I confirm that spelling influences my total score and I should choose Turkish special characters (ç,ı,ğ,ö,ş,ü) when necessary

☐ I confirm that I will not use a dictionary or any other aids in completing the test

☐ I confirm that I will complete the participant feedback survey at the end.

Turkish C-test

Text 1 - Mahallem

Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir süpermar_____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

Text 2 – Danielle Clausen

Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis_____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

Text 3 - Istanbul

İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san_____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____, sinema ve kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

Text 4 - Motivasyon

Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında motiva___ _____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

Text 5 – Koku ve Tat

Koku alma duyunuz tat duyunuz ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

Text 6 – Kültürel Mekân

Mekânın, kültürel süreklilik açısından gerekli olduğu gerçeği göz önüne alındığında, diğer alanlarda olduğu gibi halk oyunları alanında da kültürün üretildiği ve aktarıldığı kültürel mekânların üstlendiği işlevin irdelenme gereği kaçınılmazdır. Bu nok_____ halk oyunl_____ yaşatıldığı ve gel_____ kuşaklara aktar_____ kültürel mekân_____ yok ol_____ çekincesi, kült_____ de yok ol_____ çekincesini berab_____ getirir. Gelen_____ temsillerde önce_____ olan mekâ_____, günümüz koşull_____ küresel ve ye_____ etkilerle değiş_____. Bu ned_____ kültürel ve mekâ_____ farklılaşma ve çeşit_____ hızlanmıştır. Bun_____ birlikte, ya_____ koşullarındaki hızlı değişim, evrensel kültür ile yerel kültürler arasındaki çelişki, kültür ve mekân etkileşiminde yeni boyutlar yaratmış ve gelenek yeniden biçimlenen bu mekânlarda yaşatılır hâle gelmiştir.

Appendix 9: Feedback Survey for L2 learners in Validation Study 1

Please answer the following questions about the study.

Section 1: The following questions relate to your test taking experience.

Q1 Was there anything that confused you while completing the test and the questionnaire?

- ○ Yes
- ○ No

Q2 Please write what confused you while completing the test and the questionnaire.

_____

Q3 Did you have any problems with logging in and navigation tools?

- ○ Yes
- ○ No

Q4 Please write what problems you had with logging in and navigation tools.

_____

Q5 Did taking the test without supervision have any impact on your performance?

- ○ Yes
- ○ No

Q6 Please write the impacts of taking the test without supervision on your performance.

_____

Section 2: The following questions relate to your views about the Turkish C-test that you have just taken. Please select the level of difficulty for each text (you don't need to fill in gaps again!).

Q7
Text 1
Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir

süpermar_____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 1	○	○	○	○	○

Q8

Text 2

Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis_____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 2	○	○	○	○	○

Q9

Text 3

İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san____ _____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslarar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____, sinema ve kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 3	○	○	○	○	○

Q10

Text 4

Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında motiva_____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran

b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 4	○	○	○	○	○

Q11

Text 5

Koku alma duyunuz tat duyunuz ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 5	○	○	○	○	○

Q12

Text 6

Mekânın, kültürel süreklilik açısından gerekli olduğu gerçeği göz önüne alındığında, diğer alanlarda olduğu gibi halk oyunları alanında da kültürün üretildiği ve aktarıldığı kültürel mekânların üstlendiği işlevin irdelenme gereği kaçınılmazdır. Bu nok_____ halk oyunl_____ yaşatıldığı ve gel_____ kuşaklara aktar_____ kültürel mekân_____ yok ol_____ çekincesi, kült_____ de yok ol_____ çekincesini berab_____ getirir. Gelen_____ temsillerde önce_____ olan mekâ_____, günümüz koşull_____ küresel ve ye_____ etkilerle deği_____. Bu ned_____ kültürel ve mekâ_____ farklılaşma ve çeşit_____ hızlanmıştır. Bun_____ birlikte, ya_____ koşullarındaki hızlı değişim, evrensel kültür ile yerel kültürler arasındaki çelişki, kültür ve mekân etkileşiminde yeni boyutlar yaratmış ve gelenek yeniden biçimlenen bu mekânlarda yaşatılır hâle gelmiştir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 6	○	○	○	○	○

Q13 If you found some texts more difficult than others, write the reasons why you found them more difficult.

_____

Q14 Among the 6 texts, have you seen any of them before?

○ Yes (1)

○ No (2)

Q15 Where have you seen them before?

_____

Q16 Please choose what you think about the following statement.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The Turkish C-test above is a good and fair estimate of Turkish language ability	○	○	○	○	○

Q17 Please explain your response for the statement above (i.e., why you chose it).

_____

Q18 If you have any other comments, please feel free to write below.

_____

Q19 Do you wish to receive a $5 (£5) Starbucks electronic gift card for your participation?

○ Yes (1)

○ No (2)

Q20 To receive the electronic Starbucks gift card, please write your e-mail address below.

_____

Appendix 10: SLA Researcher Survey for Validation Study 1

SLA Researcher Survey

Please fill in this form about your background and your views of the Turkish C-test.

Section 1: The following questions relate to your general background.

Q1 What is your primary role at your institution?

- ○ Graduate Student
- ○ Research/Teaching Assistant
- ○ Postdoctoral Researcher
- ○ Professor
- ○ Other (please specify)

_____

Q2 What is your gender?

- ○ Male
- ○ Female
- ○ Other
- ○ Prefer not to say

Q3 Please write your age.

_____

Q4 Which country are you working/studying in?

_____

Q5 What is your mother tongue/first language? (eg. Turkish)

_____

Q6 Are you a heritage speaker of Turkish? (a person raised in a home where a non-majority language (eg. Turkish) is spoken is a **heritage speaker** of that language if she/he possesses some proficiency in it)

- ○ Yes
- ○ No

Q7 Are you a bilingual speaker of Turkish? (**bilingual speaker** means a person who has learned two or more languages relatively simultaneously during early childhood)

- ○ Yes
- ○ No

Section 2: The following questions relate your experience and views regarding C-tests.

Q8 Have you ever read about C-tests before? (C-test is an alternative to cloze test where some letters are deleted rather than words)

- ○ Yes
- ○ No

Q9 Have you ever used a C-test before?

- ○ Yes
- ○ No

Q10 What was the language of the C-test that you have used?

_____

Q11 Here is an overview and example of C-tests. Please read it and state how much you agree with the statements below.

The C-test is a type of reduced redundancy test which provides short-cut estimates of overall language proficiency. It typically consists of four to six short texts with 20 or 25 gaps in each. In these short texts, second half of each second word is deleted starting from the second sentence. The first and last sentences are left intact. The C-test is very practical given the ease of development, administration, and scoring in a short amount of time. It can be completed within 20 minutes for a typical 4-text C-test (i.e., 5 minutes per text). This is an example C-test passage. Starting wi_____ the sec____ word o____ this sent____, the la____ half fr____ each consec____ word h____ been del____. C-tests a____ typically comp____ of mult____ texts wh____ become increa____ difficult. Ea____ text usu____ contains bet____ 20 a____ 25 it____. So, wh____ do y____ think ge____ tested on a C-test? Let's examine C-tests more closely

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The C-test overview and example above provided enough information about the test format	○	○	○	○	○
C-tests are good and fair estimates of language ability.	○	○	○	○	○
C-tests are useful to quickly estimate language learners' overall language proficiency levels	○	○	○	○	○

Section 3: The remaining questions are about your views regarding a Turkish C-test.

Q12 Here is the online Turkish C-test instruction. Please read it and state how much you agree with the statements below.

In the following 6 texts, some letters are missing in a number of words. Please fill in

the gaps by completing these words as shown in the example below.

Example:
Geçen Paz_____ önemli işle_____ bitirdim ve ta_____ için pl_____ yaptım.
Geçen Pazar önemli işlerimi bitirdim ve tatil için plan yaptım.
Don't panic! You may not be able to fill in all of the gaps; but it is very important that you try your best to complete all the blanks in all texts.

You can choose Turkish special characters (ç, ı, ğ, ö, ü, ş) from the text box above the word that you are completing if necessary. Spelling counts, so be as accurate as possible. Pay attention to context, vocabulary, and grammar.

You have a maximum of 30 minutes to fill in all the gaps in all texts (approximately 5 minutes per text). If you haven't completed the test by the time limit, the test will be submitted automatically. So, it is recommended that you spend about 5 minutes per text. Your remaining time will be displayed on the screen.

After you fill in the gaps in each text, read over the text to check there are no typos and make sure your answers are consistent with the rest of the text such that you use appropriate verb tenses and personal pronoun markers.

Your estimated score will be displayed at the end of the test. Do not forget to complete the short participant feedback survey at the end.

Do not use a dictionary or any other aids in completing the test.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The Turkish C-test example was clear.	○	○	○	○	○
The Turkish C-test instructions provided enough information about the test format.	○	○	○	○	○

Please review the Turkish C-test texts below and indicate the level of difficulty for each text from the perspective of Turkish foreign language learners (i.e., how difficult text 1 is for Turkish language learners). You are not expected to take the test.

Q13

Text 1

Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir süpermar_____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 1	○	○	○	○	○

Q14

Text 2

Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis_____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 2	○	○	○	○	○

Q15

Text 3

İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san_____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslarar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____, sinema ve

kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 3	○	○	○	○	○

Q16
Text 4
Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında motiva_____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 4	○	○	○	○	○

Q17
Text 5
Koku alma duyunuz tat duyunuz ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 5	○	○	○	○	○

Q18
Text 6
Mekânın, kültürel süreklilik açısından gerekli olduğu gerçeği göz önüne alındığında, diğer alanlarda olduğu gibi halk oyunları alanında da kültürün üretildiği ve aktarıldığı kültürel mekânların üstlendiği işlevin irdelenme gereği kaçınılmazdır. Bu nok_____ halk oyunl_____ yaşatıldığı ve gel_____ kuşaklara

aktar_____ kültürel mekân_____ yok ol_____ çekincesi, kült_____ de yok ol_____ çekincesini berab_____ getirir. Gelen_____ temsillerde önce_____ olan mekâ_____, günümüz koşull_____ küresel ve ye_____ etkilerle değiş_____. Bu ned_____ kültürel ve mekâ_____ farklılaşma ve çeşit_____ hızlanmıştır. Bun_____ birlikte, ya_____ koşullarındaki hızlı değişim, evrensel kültür ile yerel kültürler arasındaki çelişki, kültür ve mekân etkileşiminde yeni boyutlar yaratmış ve gelenek yeniden biçimlenen bu mekânlarda yaşatılır hâle gelmiştir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 6	○	○	○	○	○

Q19 If you found some texts more difficult than others, write the reasons why you found them more difficult.

Q20 Please choose your level of agreement for the two statements below.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The Turkish C-test above is a good and fair estimate of Turkish language ability	○	○	○	○	○
The Turkish C-test above will be useful in my research studies to quickly estimate my participants' overall Turkish language proficiency levels	○	○	○	○	○

Q21 Please explain your responses for the two statements above (i.e., why you chose them).

_____

Q22 If you have any other comments or suggestions about the Turkish C-test, please feel free to write them below.

_____

Q23 Do you wish to receive a £5 ($5) electronic Starbucks gift card for your participation?

- ○ Yes
- ○ No

Q24 To receive the electronic gift card, please write your e-mail address below.

_____

Q25 Do you wish to participate in a follow-up interview about your survey responses through Skype or Zoom and enter a £50 ($50) Amazon prize draw?

- ○ Yes
- ○ No

Q26 Would you prefer a video call or a phone call on Skype/Zoom?

- ○ Video Call
- ○ Phone Call
- ○ Either is fine

Q27 To be contacted for the interview, please write your e-mail address below

_____

Q28 Please book a Skype interview date on my calendar below (it takes a little time for the calendar to appear, please wait!) To do this:

1) choose your time zone
2) pick up a time slot and confirm it
3) write your contact details
4) press on 'schedule event'
Then, **DON'T FORGET** to click on the next button for the end of the survey!

Skype Interview

30 min

Thank you for agreeing to participate in a follow-up interview! The interview will take take a maximum of 30 min and you will participate in a £50 Amazon prize draw.

Select a Date & Time

September 2019

SUN	MON	TUE	WED	THU	FRI	SAT
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

Appendix 11: Interview Questions for SLA Researchers

1. Can you tell me a little bit about yourself?
    a) What's your educational and professional background?
    b) What kind of research have you done?

2. What do you think about language assessment in SLA research?
    c) How often do you need to estimate your participants' proficiency levels?
    d) Do you find it easy to reach validated Turkish proficiency measurements?

3. What is your impression of C-tests in general?

4. What is your impression of the Turkish C-test?
    e) What do you think of the given Turkish C-test in terms of its difficulty, usefulness, appropriateness and fairness for Turkish L2 learners?
    f) In which contexts do you think Turkish C-test would function better (i.e., diagnostic test, placement test)? Why?

5. Would you use the Turkish C-test as an estimate of overall language proficiency in your research studies? Why or why not?

Appendix 12: Student Information Sheet and Consent Form

Dear Participant,

My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am doing the evaluation of a Turkish test for foreign language learners of Turkish.

I invite you to participate in my study which consists of three parts: **1) Background Questionnaire, 2) Turkish C-test, 3) Participant Feedback Survey**.

The study should take around **30 minutes** in total. You will receive **a £5 (or $5) Starbucks electronic gift card** when you complete all three parts of the study.

Please don't take the test on mobile devices since it is not mobile-friendly.

Participation in this study is anonymous and your responses will be kept completely confidential. Please read the information sheet below before you participate in the study.

Thank you for your interest.

If you agree to participate, please tick the boxes below:

□ I confirm that I have read and understood the information sheet of the study above.

□ I agree to take part in the study above.

Participant Information Sheet

Project Title: Validating Language Tests in Second Language Acquisition Research and Educational Programs Through an Argument-Based Approach: The C-test in Turkish

Researcher: Merve Demiralp (University of Bristol)

You are invited to participate in this research study. Please read the following information before you decide whether you wish to participate or not.

What is the aim of this study?
The aim of this study is to evaluate a Turkish language test as an estimate of overall language proficiency in research settings.

Why have I been invited?
You are being asked to take part in this study because you have been learning & have learned Turkish as a second/foreign language in academic settings.

What will I be asked to do if I take part in this study?
If you agree to participate, you will be asked to take a background questionnaire, a Turkish C-test, and a participant feedback survey. The Turkish C-Test consists of

short Turkish reading passages in which some letters are missing in a number of words. You will try to fill in the missing letters. The study should last around 30 minutes in total.

What are the possible benefits of taking part?

If you complete the test together with questionnaires, you will receive a £5 (or $5) Starbucks gift card through the email address you provide upon your completion of the study. Also, you will practice your language skills in Turkish. Information collected in this study may benefit language researchers, teachers and testers in the future.

What are the possible risks and advantages of taking part?

There are no anticipates risks associated with this study.

What will happen if I decide not to take part or not to carry on with this study?

Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw at any time by logging out during the test. However, if you withdraw, you will not get the rewards associated with the study and your test will not be used in this study. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

Will my taking part in this study be kept confidential?

What you write during this study will remain anonymous and cannot be linked to you in any way. Study data will be protected in the researcher's personal and password protected laptop. Only the researcher will have access to the data.

What will happen to the results of the study?

The results of the study will only be used for academic purposes. The study will be a part of the researcher's dissertation study. It can also be used in academic conferences and publications.

What if there is a problem?

If you have any questions or complaints regarding this research project in general, please feel free to contact the researcher at md15381@bristol.ac.uk or her supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk. Now, if you agree to participate, please tick the consent buttons on the previous page.

Appendix 13: Researcher Information Sheet and Consent Form

Dear Participant,

My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am investigating researchers' perception of a Turkish test for foreign language learners of Turkish.

I invite you to participate in my study and take the following survey, which will take less than 30 minutes. You will receive **a £5 (or $5) Starbucks electronic gift card** if you complete the survey and provide your e-mail address.

Participation in this study is anonymous and your responses will be kept completely confidential. Please read the information sheet below before you participate in the study.

Thank you for your interest.

If you agree to participate, please tick the boxes below:

□ I confirm that I have read and understood the information sheet of the study above.

□ I agree to take part in the study above.

Participant Information Sheet

Project Title: Validating Language Tests in Second Language Acquisition Research and Educational Programs through an Argument-based Approach: The C-test in Turkish

Researcher: Merve Demiralp (University of Bristol)

You are invited to participate in this research study. Please read the following information before you decide whether you wish to participate or not.

What is the aim of this study?
The aim of this study is to evaluate a Turkish language test as an estimate of overall language proficiency.

Why have I been invited?
You are being asked to take part in this study because you have been doing research in Turkish as a second language.

What will I be asked to do if I take part in this study?
If you agree to participate, you will be asked to take a short survey about your demographic information and your views about the Turkish C-test. The survey should last less than 30 minutes.

What are the possible benefits of taking part?
If you complete the survey, you will receive a £5 (or $5) Starbucks gift card through

the email address you provide upon your completion of the study. Information collected in this study may benefit language researchers, teachers and testers in the future.

What are the possible risks of taking part?
There are no anticipated risks associated with this study.

What will happen if I decide not to take part or not to carry on with this study?
Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw at any time by logging out during the survey. However, if you withdraw, you will not get the rewards associated with the study and your data will not be used in this study. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

Will my taking part in this study be kept confidential?
What you write during this study will remain anonymous and cannot be linked to you in any way. Study data will be protected in the researcher's personal and password protected laptop. Only the researcher will have access to the data.

What will happen to the results of the study?
The results of the study will only be used for academic purposes. The study will be a part of the researcher's dissertation study. It can also be used in academic conferences and publications.

What if there is a problem?
If you have any questions or complaints regarding this research project in general, please feel free to contact the researcher at md15381@bristol.ac.uk or her supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk. Now, if you agree to participate, please tick the consent buttons on the previous page.

Appendix 14: Researcher Interviewee Information Sheet and Consent Form

Dear Participant,
My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am looking at researchers' perception of a Turkish language test that I have developed. You are being asked to take part in this study because you are a researcher who has been working on Turkish as a second language in academic settings. Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw the study at any time by letting me know. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

If you agree to participate, you will be asked to be interviewed about the survey that you have taken before. The interview will be done through Skype or Zoom, and it will be recorded. The transcript of the interview will be sent to you afterwards. The interview will involve questions about your research experience with learners of Turkish and views of the Turkish C-test. It will take around **30 minutes.** What you say during the interview will remain anonymous. I will not use your names or any special information about you and you will not be identifiable in my thesis or any published material. Study data will be kept in my personal and password protected laptop. Access to the study data will be protected. Only I will have access to the data.

There are no risks associated with this study. If you participate in the interview, you will also enter a **£50 Prize Draw (Amazon gift card)**. Information collected in this study may benefit language researchers, teachers and testers in the future.

If you have any questions or complaints regarding this research project in general, please feel free to contact me at md15381@bristol.ac.uk or my supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk. Thank you for your interest. *This project has been approved by the Graduate School of Education's Research Ethics Committee at the University of Bristol.*

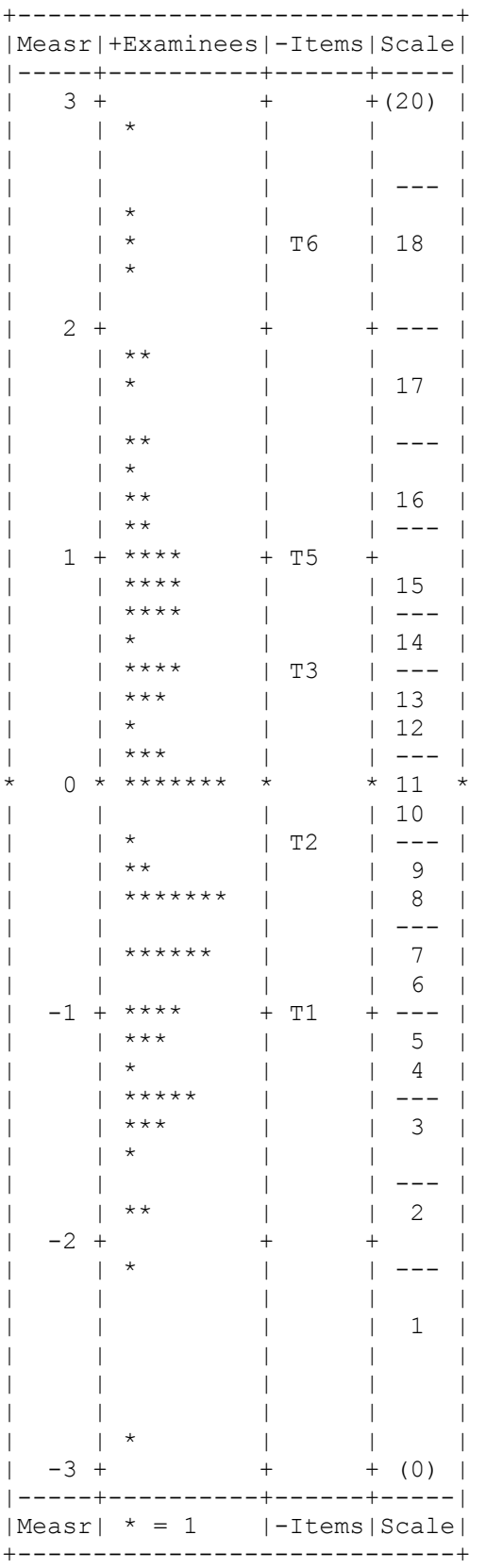
If you agree to participate, please tick the boxes below:

☐ I confirm that I have read and understood the information sheet of the study above.

☐ I consent that the interview will be recorded.

☐ I agree to take part in the study above.

Appendix 15: Rasch Analysis with 82 Examinees in Validation Study 1



Variance explained by Rasch model = 93.07%
Separation = 4.33; Strata = 6.11; Reliability = .95

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.77	.85	.92	1.21	.06	-.96
T3	.88	1.32	.60	.67	.06	-.24
T4	.91	1.32	.68	.67	.05	.48
T9	.85	.89	.88	1.08	.06	1.00
T12	.74	.89	.93	.98	.08	2.41

Appendix 16: DIF Measure

PERSON CLASS/	Obs-Exp Average	DIF MEASURE	DIF S.E.	PERSON CLASS/	Obs-Exp Average	DIF MEASURE	DIF S.E.	DIF CONTRAST	JOINT S.E.	Rasch-Welch t	d.f.	Prob.	Mantel Chi-squ	Prob.	Size CUMLOR	Active Slices	ITEM Number	Name
UK	.22	-1.58	.09	USA	-.14	-1.48	.08	-.11	.12	-.87	68	.3884	.0037	.9515	.05	11	1	Text 1
UK	.29	-.89	.09	USA	-.18	-.77	.07	-.12	.11	-1.06	68	.2939	.0069	.9339	-.06	11	2	Text 3
UK	-.21	-.01	.09	USA	.13	-.10	.07	.09	.11	.76	63	.4491	.6560	.4180	-.55	11	3	Text 4
UK	-.01	.49	.10	USA	.02	.49	.07	.00	.12	.00	61	1.000	.0010	.9748	-.02	11	4	Text 9
UK	-.27	2.08	.14	USA	.18	1.83	.09	.25	.17	1.47	54	.1480	1.0857	.2974	1.04	11	5	Text 12
USA	-.14	-1.48	.08	UK	.22	-1.58	.09	.11	.12	.87	68	.3884	.0037	.9515	-.05	11	1	Text 1
USA	-.18	-.77	.07	UK	.29	-.89	.09	.12	.11	1.06	68	.2939	.0069	.9339	.06	11	2	Text 3
USA	.13	-.10	.07	UK	-.21	-.01	.09	-.09	.11	-.76	63	.4491	.6560	.4180	.55	11	3	Text 4
USA	.02	.49	.07	UK	-.01	.49	.10	.00	.12	.00	61	1.000	.0010	.9748	.02	11	4	Text 9
USA	.18	1.83	.09	UK	-.27	2.08	.14	-.25	.17	-1.47	54	.1480	1.0857	.2974	-1.04	11	5	Text 12

Width of Mantel slice: MHSLICE = .010 logits, Zero cell adjustment: MHZERO = .0000

Appendix 17: Spearman's Rho Correlations with 6-text C-test in Validation Study 1

	Turkish C-Test Scores (6-Text)
months of study	.56
months in Turkey	.51
age of learning	-.46
hours of study per week	.25
institutional level	77
self-reading	.81
self-writing	.82
self-speaking	.82
self-listening	.80
self-overall	.82

Appendix 18: Background Questionnaire for TYS Candidates in Validation Study 2

Turkish Language Learning Background Questionnaire

Before you take the test, please fill out the following background questionnaire.

Section 1: The following questions relate to your demographic information.

Q1 Please write your age.

_____

Q2 What is your gender?

- ○ Male
- ○ Female
- ○ Other
- ○ Prefer not to say

Q3 Which country are you currently working/studying in?

_____

Q4 Please write your mother tongue/first language (eg. English).

_____

Q5 Are you a heritage speaker of Turkish? (a person raised in a home where a non-majority language (eg. Turkish) is spoken is a **heritage speaker** of that language if she/he possesses some proficiency in it)

- ○ Yes
- ○ No

Q6 Are you a bilingual speaker Turkish? (**bilingual speaker** means a person who has learned two or more languages relatively simultaneously during early childhood)

- ○ Yes
- ○ No

Q7 What is your highest completed degree of education?

- ○ High School
- ○ Undergraduate Degree
- ○ Master's Degree
- ○ PhD

Q8 What's your subject of study (or degree program/ discipline/ job)?

_____

Q9 Do you have any learning difficulties (eg. dyslexia) that can cause problems with reading, writing or spelling?

- ○ Yes
- ○ No

Q10 What is the type of your learning difficulty?

_____

Section 2: The following questions relate to your Turkish language learning experience.

Q11 Please list the names and choose the **levels** of Turkish language classes (i.e., beginner, intermediate, advanced) that you have taken and you are currently taking. Indicate if they were/are required and when & where you took/are taking them.

	Level	When?	Where?	Required
Course Name				
Course Name				
Course Name				
Course Name				
Course Name				

Q12 At what age did you begin studying Turkish?

_____

Q13 How long have you been studying Turkish? (in years and months)

_____

Q14 Have you ever visited a Turkish speaking country?

- ○ Yes
- ○ No

Q15 Please indicate where and how long you visited.

Location	Length of visit

Q16 Outside of class, how many hours per week do you spend using Turkish?

_____

Q17 Do you have family members who speak Turkish?

- ○ Yes
- ○ No

Q18 Please indicate the family members who speak Turkish (e.g., parents, grandparents, etc.).

_____

Section 3: The following question relates to your learning experience with other languages.

Q19 Please list any other languages that you have previously studied and the length of your study in years and months.

Language	Length of study

Section 4: The following questions relate your Turkish language proficiency. They are crucial information and cannot be left blank.

Q20 Please write the date and location you took the Turkish Proficiency Exam (TYS) organised by the Yunus Emre Institution. (i.e, July 2018 - London)

_____

Q21 Please write **your total score, your level** (i.e., below B2, B2, C1, C2), and **your scores in four sections of the Turkish Proficiency Exam** (reading, listening, writing, speaking).

	Total Score	Level	Reading	Listening	Writing	Speaking
TYS						

Q22 Please circle your self-perceived proficiency level for Turkish in the following areas.

	Beginner	**Elementary**	**Intermediate**	**Advanced**	**Very advanced**
Reading	○	○	○	○	○
Writing	○	○	○	○	○
Listening	○	○	○	○	○
Speaking	○	○	○	○	○
Overall Proficiency	○	○	○	○	○

Section 5: The remaining question relates to your experience with computers.

Q23 Please circle how comfortable you are with computer in the given conditions.

	Extremely comfortable	Moderately comfortable	Somewhat comfortable	A little bit comfortable	Not at all comfortable
using a computer	○	○	○	○	○
taking a test through a computer	○	○	○	○	○

Thank you for completing the questionnaire! *Now, please click on the next button to see the test instructions.*

Appendix 19: Turkish C-test for Validation Study 2

Instructions for the test

Please read carefully. The test has not started yet.

In the following 8 texts, some letters are missing in a number of words. Please fill in the gaps by completing these words as shown in the example below.

Example:
Geçen Paz_____ önemli işle_____ bitirdim ve ta_____ için pl_____ yaptım.
Geçen Paz**ar** önemli işle**rimi** bitirdim ve ta**til** için pl**an** yaptım.

Don't panic! You may not be able to fill in all of the gaps; **but it is very important that you try your best to complete all the blanks in all texts.**

You can choose Turkish special characters (ç, ı, ğ, ö, ü, ş) from the text box above the word that you are completing if necessary. **Spelling counts,** so be as accurate as possible. Pay attention to context, vocabulary, and grammar.

You have a maximum of **45 minutes** to fill in all the gaps in all texts. If you haven't completed the test by the time limit, the test will be submitted automatically. So, it's recommended that you spend about 5 minutes per text. Your remaining time will be displayed on the screen. Please try to use all the time and make your best effort.

You will be directed to the next text as soon as you have clicked on "submit" and "next". **You cannot go back once you click submit to a given text**.

After you fill in the gaps in each text, **read over the text to check there are no typos and make sure your answers are consistent with the rest of the text** such that you use appropriate verb tenses and personal pronoun markers. Your estimated score will be displayed at the end of the test. **Do not forget to complete the short participant feedback survey at the end.** Do **not use a dictionary or any other aids** in completing the test. Now, please tick the boxes below and click on "START THE TEST" button when you are ready.

☐ I confirm that I cannot go back to the texts that I have submitted

☐ I confirm that spelling influences my total score and I should choose Turkish special characters (ç,ı,ğ,ö,ş,ü) when necessary

☐ I confirm that I will not use a dictionary or any other aids in completing the test

☐ I confirm that I will complete the participant feedback survey at the end.

Turkish C-test

Text 1 - Mahallem

Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir süpermar_____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

Text 2 – Danielle Clausen

Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis_____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

Text 3 - Istanbul

İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san_____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____, sinema ve kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

Text 4 – Anakent Koleji

Anakent Koleji, öğrencilerini geleceğe tam olarak hazırlamayı misyon edinen bir kurumdur. Okulumuzda yab_____ dil öğret_____ büyük ön_____ verilir. Öğren_____ İngilizce ve Alm_____ dillerini sı_____ ortamında ve aktiv_____ yoluyla öğren_____. Bu sür_____, hazırlık sınıfl_____ başlayarak 12. sın_____ kadar de_____ eder. D_____ öğrenme konus_____ en öne_____ etkinliğimiz Yab_____ Diller Kulü_____. Bu ku_____ küresel kon_____ hakkında uluslar_____ çalışmalar yapar. Yabancı bir dili etkin bir şekilde kullanma adına kulüp çalışmalarımız önem taşır.

Text 5 - Motivasyon

Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında motiva_____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

Text 6 – Koku ve Tat

Koku alma duyunuz tat duyunuz ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

Text 7 – Bilim Kadınları

Türkiye'deki bilim kadınları hakkında yapılan hemen her araştırma kimi ilginç olguların altını çizer. İlk ola_____, üniversitelerin far_____ kademelerinde y_____ alan kadın_____ oranı s_____ derece yüks_____. Sadece öğr_____ ya da asis_____ düzeyinde değ_____, öğretim üy_____ ve yöne__ _____ kadrolarındaki kadın_____ oranı da b_____ hayli kabar_____. Bununla ber_____, bilim kadı_____, kadın olma_____ dolayı he_____ hemen hiçb_____ ayrımcılığa uğramad_____ dile getirmektedirler. Bu saptamayı takip ettiğimizde 1930'lardan bu yana bir süreklilik buluruz.

Text 8 - Kültürel Mekân

Mekânın, kültürel süreklilik açısından gerekli olduğu gerçeği göz önüne alındığında, diğer alanlarda olduğu gibi halk oyunları alanında da kültürün üretildiği ve aktarıldığı kültürel mekânların üstlendiği işlevin irdelenme gereği kaçınılmazdır. Bu nok_____ halk oyunl_____ yaşatıldığı ve gel_____ kuşaklara aktar_____ kültürel mekân_____ yok ol_____ çekincesi, kült_____ de yok ol_____ çekincesini berab_____ getirir. Gelen_____ temsillerde önce_____ olan mekâ_____, günümüz koşull_____ küresel ve ye_____ etkilerle değiş_____. Bu ned_____ kültürel ve mekâ_____ farklılaşma ve çeşit_____ hızlanmıştır. Bun_____ birlikte, ya_____ koşullarındaki hızlı değişim, evrensel kültür ile yerel kültürler arasındaki çelişki, kültür ve mekân etkileşiminde yeni boyutlar yaratmış ve gelenek yeniden biçimlenen bu mekânlarda yaşatılır hâle gelmiştir.

Thank you! Please click on the next page button to complete the survey.

Appendix 20: Feedback Survey for L2 learners in Validation Study 1

Please answer the following questions about the study.

Section 1: The following questions relate to your test taking experience.

Q1 Was there anything that confused you while completing the test and the questionnaire?

- ○ Yes
- ○ No

Q2 Please write what confused you while completing the test and the questionnaire.

_____

Q3 Did you have any problems with logging in and navigation tools?

- ○ Yes
- ○ No

Q4 Please write what problems you had with logging in and navigation tools.

_____

Q5 Did taking the test without supervision have any impact on your performance?

- ○ Yes
- ○ No

Q6 Please write the impacts of taking the test without supervision on your performance.

_____

Section 2: The following questions relate to your views about the Turkish C-test that you have just taken. Please select the level of difficulty for each text (you don't need to fill in gaps again!).

Q7
Text 1
Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal

v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir süpermar_____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 1	○	○	○	○	○

Q8

Text 2

Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis_____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 2	○	○	○	○	○

Q9

Text 3

İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san_____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslarar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____, sinema ve kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 3	○	○	○	○	○

Q10

Text 4

Anakent Koleji, öğrencilerini geleceğe tam olarak hazırlamayı misyon edinen bir kurumdur. Okulumuzda yab_____ dil öğret_____ büyük ön_____ verilir. Öğren_____ İngilizce ve Alm_____ dillerini sı_____ ortamında ve aktiv_____ yoluyla öğren_____. Bu sür_____, hazırlık

sınıfl_____ başlayarak 12. sın_____ kadar de_____ eder. D_____ öğrenme konus_____ en öne_____ etkinliğimiz Yab_____ Diller Kulü_____. Bu ku_____ küresel kon_____ hakkında uluslarar_____ çalışmalar yapar. Yabancı bir dili etkin bir şekilde kullanma adına kulüp çalışmalarımız önem taşır.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 4	○	○	○	○	○

Q11

Text 5

Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında motiva_____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 5	○	○	○	○	○

Q12

Text 6

Koku alma duyunuz tat duyunuz ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 6	○	○	○	○	○

Q13

Text 7

Türkiye'deki bilim kadınları hakkında yapılan hemen her araştırma kimi ilginç olguların altını çizer. İlk ola_____, üniversitelerin far_____ kademelerinde y_____ alan kadın_____ oranı s_____ derece yüks_____. Sadece öğr_____ ya da asis_____ düzeyinde değ_____, öğretim

üy_____ ve yöne_____ kadrolarındaki kadın_____ oranı da b_____ hayli kabar_____. Bununla ber_____, bilim kadı_____, kadın olma_____ dolayı he_____ hemen hiçb_____ ayrımcılığa uğramad_____ dile getirmektedirler. Bu saptamayı takip ettiğimizde 1930'lardan bu yana bir süreklilik buluruz.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 7	○	○	○	○	○

Q14
Text 8
Mekânın, kültürel süreklilik açısından gerekli olduğu gerçeği göz önüne alındığında, diğer alanlarda olduğu gibi halk oyunları alanında da kültürün üretildiği ve aktarıldığı kültürel mekânların üstlendiği işlevin irdelenme gereği kaçınılmazdır. Bu nok_____ halk oyunl_____ yaşatıldığı ve gel_____ kuşaklara aktar_____ kültürel mekân_____ yok ol_____ çekincesi, kült_____ de yok ol_____ çekincesini berab_____ getirir. Gelen_____ temsillerde önce_____ olan mekâ_____, günümüz koşull_____ küresel ve ye_____ etkilerle deği_____. Bu ned_____ kültürel ve mekâ_____ farklılaşma ve çeşit_____ hızlanmıştır. Bun_____ birlikte, ya_____ koşullarındaki hızlı değişim, evrensel kültür ile yerel kültürler arasındaki çelişki, kültür ve mekân etkileşiminde yeni boyutlar yaratmış ve gelenek yeniden biçimlenen bu mekânlarda yaşatılır hâle gelmiştir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 8	○	○	○	○	○

Q15 If you found some texts more difficult than others, write the reasons why you found them more difficult.

_____

Q16 Among the 8 texts, have you seen any of them before?

○ Yes

○ No

Q17 Where have you seen them before?

_____

Q18 Please choose what you think about the following statement.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The Turkish C-test above is a good and fair estimate of Turkish language ability	○	○	○	○	○
The Turkish C-test above will be useful to practice my language skills before taking TYS.	○	○	○	○	○
The Turkish C-test above will be useful to quickly estimate my TYS level	○	○	○	○	○

Q19 Please explain your response for the statement above (i.e., why you chose it).

_____

Q20 If you have any other comments, please feel free to write below.

_____

Q21 Do you wish to receive a $5 (£5) Starbucks electronic gift card or 25 TL Idefix electronic gift card for your participation? Please choose your response below.

○ Yes, I want to receive $5 Starbucks electronic gift card

○ Yes, I want to receive £5 Starbucks electronic card

○ Yes, I want to receive 25 TL Idefix electronic gift card

○ No, I don't want to receive any gift cards.

Q22 To receive the electronic Starbucks gift card, please write your e-mail address below.

_____

Q23 Do you wish to participate in a follow-up interview about your feedback through Skype or Zoom and enter a £50 Prize Draw? (The interview can be done in English or Turkish)

- ○ Yes
- ○ No

Q24 Would you prefer a video call or a phone call on Skype/Zoom?

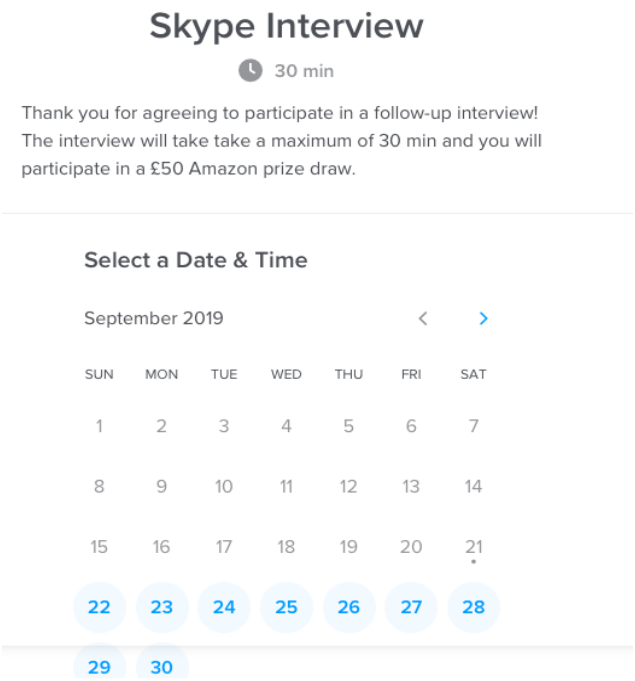
- ○ Video Call
- ○ Phone Call
- ○ Either is fine

Q25 To be contacted for the interview, please write your e-mail address below.

_____

Q26 Please book a Skype interview date on my calendar below (it takes a little time for the calendar to appear, please wait!) To do this:

1) choose your time zone
2) pick up a time slot and confirm it
3) write your contact details
4) press on 'schedule event'
Then, DON'T FORGET to click on the next button for the end of the survey!



Appendix 21: Instructor Survey for Validation Study 2

Instructor Survey

Please fill in this form about your background and your views of the Turkish C-test as well as the Turkish Proficiency Exam.

Section 1: The following questions relate to your general background.

Q1 Please choose your primary role(s) at your institution.

- ☐ Language teacher
- ☐ Test developer
- ☐ Professional examiner/rater
- ☐ Curriculum coordinator
- ☐ Other (please specify) _____

Q2 What is your gender?

- ○ Male
- ○ Female
- ○ Other
- ○ Prefer not to say

Q3 Please write your age.

_____

Q4 Which country are you working in?

_____

Q5 What is your mother tongue/first language? (eg. Turkish)

_____

Q6 Are you a heritage speaker of Turkish? (a person raised in a home where a non-majority language (eg. Turkish) is spoken is a heritage speaker of that language if she/he possesses some proficiency in it)

- ○ Yes
- ○ No

Q7 Are you a bilingual speaker of Turkish? (**bilingual speaker** means a person who has learned two or more languages relatively simultaneously during early childhood)

- ○ Yes
- ○ No

Section 2: The following questions relate to the Turkish Proficiency Exam (TYS).

Q8 Have you ever read or heard about the Turkish Proficiency Exam (TYS) conducted by Yunus Emre Institution?

- ○ Yes (1)
- ○ No (2)

Q9 Have you ever used TYS scores before?

- ○ Yes
- ○ No

Q10 Here is an overview of TPE. Please read it and state how much you agree with the statements below.

TYS is developed according to the Common European Framework of Reference for Languages (CEFR), with the purpose of assessing the language proficiency of individuals learning Turkish as a second or native language. TYS consists of four sections: listening, reading, writing, and speaking. Listening and reading questions involve matching task, fill-in the gaps task, true/false questions and multiple-choice questions. Writing section involves a guided writing task and an argument essay task based on a given topic. Speaking section comprises an independent long turn speaking task and a dialogue. TYS is evaluated over 100 points with each section contributing 25 points to the total score. If the candidates achieve a minimum score of 55 in total with at least a score of 12.5 in each section, they are given the Certificate of Turkish Proficiency. The Certificate of Turkish Proficiency has three different levels according to CEFR: B2 for 55-70 points, C1 for 71-88 points, and C2 for 89-100 points. Below B2 level, no attempt is made to differentiate students and no certificate is given. Therefore, students should be at least B2 level to be successful in TYS. Below is a

table of details about TYS sections.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The TYS overview provided enough information about the test format and scoring	○	○	○	○	○
TYS is a good and fair estimate of Turkish language ability	○	○	○	○	○
TYS is useful to determine international students' admission to Turkish-medium universities	○	○	○	○	○

Section 3: The remaining questions relate to your experience and views regarding C-tests.

Q11 Have you ever read about C-tests before? (C-test is an alternative to cloze test where some letters are deleted rather than words).

○ Yes

○ No

Q12 Have you ever used a C-test before?

○ Yes

○ No

Q13 What was the language of the C-test that you have used?

_____

Q14 Here is an overview and example of C-tests. Please read it and state how much you agree with the statements below.

The C-test is a type of reduced redundancy test which provides short-cut estimates of overall language proficiency. It typically consists of four to six short texts with 20 or 25 gaps in each. In these short texts, second half of each second word is deleted starting from the second sentence. The first and last sentences are left intact. The C-test is very practical given the ease of development, administration, and scoring in a short amount of time. It can be completed within 20 minutes for a typical 4-text C-test (i.e., 5 minutes per text). This is an example C-

test passage. Starting wi_____ the sec____ word o____ this sent____, the la____ half fr____ each consec____ word h____ been del____. C-tests a____ typically comp____ of mult____ texts wh____ become increa____ difficult. Ea____ text usu____ contains bet____ 20 a____ 25 it____ . So, wh____ do y____ think ge____ tested on a C-test? Let's examine C-tests more closely

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The C-test overview and example above provided enough information about the test format	○	○	○	○	○
C-tests are good and fair estimates of language ability.	○	○	○	○	○
C-tests are useful to quickly estimate language learners' overall language proficiency levels	○	○	○	○	○

Section 4: The remaining questions are about your views regarding the Turkish C-test.

Q15 Here is the online Turkish C-test instruction. Please read it and state how much you agree with the statements below.

In the following 8 texts, some letters are missing in a number of words. Please fill in the gaps by completing these words as shown in the example below.

Example:
Geçen Paz_____ önemli işle_____ bitirdim ve ta_____ için pl_____ yaptım.
Geçen Paz**ar** önemli işle**rimi** bitirdim ve ta**til** için pl**an** yaptım.

Don't panic! You may not be able to fill in all of the gaps; but it is very important that you try your best to complete all the blanks in all texts.

You can choose Turkish special characters (ç, ı, ğ, ö, ü, ş) from the text box above the word that you are completing if necessary. **Spelling counts**, so be as accurate as possible. Pay attention to context, vocabulary, and grammar.

You have a maximum of **45 minutes** to fill in all the gaps in all texts (approximately 5 minutes per text). If you haven't completed the test by the time limit, the test will be submitted automatically. So, it is recommended that you spend about 5 minutes per text. Your remaining time will be displayed on the screen.

After you fill in the gaps in each text, **read over the text to check there are no typos and make sure your answers are consistent with the rest of the text** such that you use appropriate verb tenses and personal pronoun markers.Your estimated score will be displayed at the end of the test. **Do not forget to complete the short participant feedback**

survey at the end. Do **not use a dictionary or any other aids in completing the test**.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The Turkish C-test example was clear	○	○	○	○	○
The Turkish C-test instructions provided enough information about the test format	○	○	○	○	○

Please **review** the Turkish C-test texts below and indicate the level of difficulty for each text from the perspective of Turkish second language learners (i.e., how difficult text 1 is for Turkish language learners). You are not expected to take the test.

Q16
Text 1
Burası benim mahallem. Benim ev_____ ana cad_____. Evimin karşı_____ bir lok_____ var. Lokan_____ servisi gü_____, ama fiya_____ biraz yük_____. İki so_____ ileride b_____ bakkal v_____, ama bü_____ değil. Ar_____ sokakta bü_____ bir süpermar_____ var. Genel_____ orada alışve_____ yaparım. Ya_____ sokakta küç_____ bir sin_____ var. Film izlemek için güzel bir yer.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 1	○	○	○	○	○

Q17
Text 2
Danielle Clausen Danimarkalı. Otuz sekiz yaş_____. Evli ve iki çoc_____ var. İki yıl_____ Türkiye'de yaş_____. İki y_____ daha kal_____ istiyor. Eş_____ Peter, Danimarka'nın Türk_____ konsolosu. Danielle de, haft_____ üç gü_____ konsoloslukta vi_____ bölümünde çalı_____. Çocukları, Anna ve Eric, öz_____ bir lis_____ okuyor. Danielle anadi_____ dışında İngi_____, Almanca ve Fran_____ konuşuyor. Türk_____ ise zor bul_____. Ama Danielle'in ak_____ bir Türkçesi var. Danielle Türkiye'de yaşamaktan çok memnun.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 2	○	○	○	○	○

Q18
Text 3
İstanbul, Türkiye'nin kuzey batısında, Avrupa ile Asya kıtaları üzerinde uzanır. Dünyada iki kıt_____ birbirine bağl_____ tek ke_____ olan İstanbul, Türkiye'nin ve Avrupa'nın e_____ kalabalık şehr_____. Kent ülk_____ kültür, san_____ ve eko_____ başkentidir. Türkiye'de bul_____ ulusal ve uluslarar_____ şirketlerin ge_____ merkezleri b_____ kentte y_____ alır. İstanbul, tar_____ ve coğ_____ konumu i_____ kozmopolit b_____ yapıya sahi_____. Birçok tiy_____, sinema ve kül_____ merkezi vardır. İstanbul'da her yıl çeşitli konserler, festivaller ve fuarlar düzenlenir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 3	○	○	○	○	○

Q19
Text 4
Anakent Koleji, öğrencilerini geleceğe tam olarak hazırlamayı misyon edinen bir kurumdur. Okulumuzda yab_____ dil öğret_____ büyük ön_____ verilir. Öğren_____ İngilizce ve Alm_____ dillerini sı_____ ortamında ve aktiv_____ yoluyla öğren_____. Bu sür_____, hazırlık sınıfl_____ başlayarak 12. sın_____ kadar de_____ eder. D_____ öğrenme konus_____ en öne_____ etkinliğimiz Yab_____ Diller Kulü_____. Bu ku_____ küresel kon_____ hakkında uluslarar_____ çalışmalar yapar. Yabancı bir dili etkin bir şekilde kullanma adına kulüp çalışmalarımız önem taşır.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 4	○	○	○	○	○

Q20
Text 5
Motivasyonu yüksek bir öğrenci derslerine daha fazla çalışır. Daha i_____ öğrenir ve da_____ başarılı ol_____. Dolayısıyla ula_____ istediği hede_____ daha hı_____ bir şek_____ ve da_____ kolay ula_____. Öğrencinin hede_____ ulaşmasında motiva_____ önemi ç_____ büyüktür. Öğret_____ sevmek de motiv_____ artıran b_____ faktördür. B_____ nedenle öğret_____ kendini öğrenc_____ sevdirmesi, onl_____ rol

mo_____ olabilmesi önemlidir. Bunu yapabilmek için verdiği sözleri tutması, öğrencileriyle iyi ilişkiler geliştirmesi gerekir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 5	○	○	○	○	○

Q21
Text 6
Koku alma duyunuz tat duyunuz ile bağlantılıdır. Yiyeceğin koku_____ alamadığınızda muhte_____ tadını da alamaz_____. Bu dur_____ yeterince ye_____ yememenize ve ki_____ kaybetmenize ne_____ olur. Vücu_____ ihtiyacı ol_____ besinleri alamad_____ için vit_____ ve min_____ eksikliği gi_____ bazı sağ_____ sorunları yaşar_____. Koku almam_____ ruh hali_____ de etkileyebilir. Çiç_____, gıda ve ben_____ kokular si_____ yaşam sevinci verir. Bu kokuları almamanız ise kendinizi üzgün veya depresif hissetmenize neden olabilir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 6	○	○	○	○	○

Q22
Text 7
Türkiye'deki bilim kadınları hakkında yapılan hemen her araştırma kimi ilginç olguların altını çizer. İlk ola_____, üniversitelerin far_____ kademelerinde y_____ alan kadın_____ oranı s_____ derece yüks_____. Sadece öğr_____ ya da asis_____ düzeyinde değ_____, öğretim üy_____ ve yöne_____ kadrolarındaki kadın_____ oranı da b_____ hayli kabar_____. Bununla ber_____, bilim kadı_____, kadın olma_____ dolayı he_____ hemen hiçb_____ ayrımcılığa uğramad_____ dile getirmektedirler. Bu saptamayı takip ettiğimizde 1930'lardan bu yana bir süreklilik buluruz

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 7	○	○	○	○	○

Q23
Text 8
Mekânın, kültürel süreklilik açısından gerekli olduğu gerçeği göz önüne alındığında, diğer alanlarda olduğu gibi halk oyunları alanında da kültürün üretildiği ve aktarıldığı kültürel mekânların üstlendiği işlevin irdelenme gereği kaçınılmazdır. Bu nok_____ halk oyunl_____ yaşatıldığı ve gel_____ kuşaklara aktar_____ kültürel mekân_____ yok ol_____ çekincesi, kült_____ de yok ol_____ çekincesini berab_____ getirir. Gelen_____ temsillerde önce_____ olan mekâ_____, günümüz koşull_____ küresel ve ye_____ etkilerle

değiş_____. Bu ned_____ kültürel ve mekâ_____ farklılaşma ve çeşit_____ hızlanmıştır. Bun_____ birlikte, ya_____ koşullarındaki hızlı değişim, evrensel kültür ile yerel kültürler arasındaki çelişki, kültür ve mekân etkileşiminde yeni boyutlar yaratmış ve gelenek yeniden biçimlenen bu mekânlarda yaşatılır hâle gelmiştir.

	Very easy	Somewhat easy	Neither easy nor difficult	Somewhat difficult	Very difficult
Text 8	○	○	○	○	○

Q24 If you found some texts more difficult than others, write the reasons why you found them more difficult.

_____

Q25 Please choose your level of agreement for the three statements below.

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The Turkish C-test above is a good and fair estimate of Turkish language ability.					
The Turkish C-test above will be useful to quickly estimate whether students are ready to take TYS (i.e., whether students are at least B2 level)					
The Turkish C-test above will be useful to estimate student levels attained by TYS (below B2, B2, C1, C2).					

Q26 Please explain your responses for the three statements above (i.e., why you chose them).

_____

Q27 If you have any other comments or suggestions about the Turkish C-test, please feel free to write them below.

Q28 Do you wish to receive a £5 ($5) Starbucks or 25 TL Idefix electronic gift card for your participation?

- ○ Yes, I want to receive £5 Starbucks e-gift card.
- ○ Yes, I want to receive $5 Starbucks e-gift card.
- ○ Yes, I want to receive 25 TL Idefix e-gift card.
- ○ No, I don't want to receive any gift cards

Q29 To receive the electronic gift card, please write your e-mail address below.

_____

Q30 Do you wish to participate in a follow-up interview about your survey responses through Skype or Zoom and enter a £50 ($50) Amazon prize draw?
Yes
No

Q31 Would you prefer a video call or a phone call on Skype/Zoom?
Video Call
Phone Call
Either is fine

Q32 To be contacted for the interview, please write your e-mail address below.
_____

Q33 Please book a Skype interview date on my calendar below (it takes a little time for the calendar to appear, please wait!) To do this:

1) choose your time zone
2) pick up a time slot and confirm it
3) write your contact details
4) press on 'schedule event'

Then, DON'T FORGET to click on the next button for the end of the survey!

Appendix 22: Interview Questions for Instructors in Validation Study 2

1. Can you tell me something about yourself?
    a) What's your educational and professional background?
    b) What kind of courses have you taught?

2. What is your impression of TYS?
    a. What are you most and least satisfied with TYS as a proficiency test?
    b. To what extent do you think TYS is a fair test?

3. What is your impression of C-tests in general?

4. What is your impression of Turkish C-test?
    a. What do you think of the given Turkish C-test in terms of its difficulty, usefulness, appropriateness and fairness for Turkish L2 learners?
    b. In which contexts do you think Turkish C-test would function better (i.e., diagnostic test, placement test)? Why?

5. Do you think Turkish C-test can be used as predictive of student performance on Turkish Proficiency Exam? Why or why not?

Appendix 23: Interview Questions for TYS Candidates in Validation Study 2

1) How is your Turkish language learning experience overall?

2) Let's talk about TYS you took earlier.
   a) Why did you take TYS?
   b) What was your impression of TYS (item format, difficulty, etc.)? How well did you do in TYS?
   c) Do you understand why you had your given level (no certificate, B2, C1, C2) on TYS? In other words, are you aware of the criterion used in TYS?
   d) Did you feel you were given the right level in TYS? Why or why not?

3) Let's talk about the Turkish C-test you took.
   a) What was your impression of Turkish C-test (item format, difficulty, etc.)? How well did you do in Turkish C-test?
   b) Do you think Turkish C-test assesses your overall language proficiency? Why or why not?
   c) How would you compare your performances in two tests?
   d) Were these performances similar, better or worse than what you expected?
   e) Did the test have any impact on you?

4) Do you think Turkish C-test can estimate TYS levels quickly?

5) Do you have some suggestions to make the given Turkish C-test a better test?

Appendix 24: Test Taker Information Sheet and Consent Form

Dear Participant,

My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am doing the evaluation of a Turkish test for foreign language learners of Turkish.

I invite you to participate in my study which consists of three parts: **1) Background Questionnaire, 2) Turkish C-test, 3) Participant Feedback Survey**.

The study should take around **45 minutes** in total. You will receive **a £5 (or $5) Starbucks or ₺ 25 (Turkish lira) Idefix (bookstore) electronic gift card** when you complete all three parts of the study.

You will be required to write your Turkish Proficiency Exam (TYS) scores and level on the questionnaire. So, please check your TYS scores before you start the study if you don't remember them.

Please don't take this test on mobile devices since it is not mobile-friendly.

Participation in this study is anonymous and your responses will be kept completely confidential. Please read this information sheet before you participate in the study.

Thank you for your interest.

If you agree to participate, please tick the boxes below:

☐ I confirm that I have read and understood the information sheet of the study above.

☐ I agree to take part in the study above.

Participant Information Sheet

Project Title: Validating Language Tests in Second Language Acquisition Research and Educational Programs through an Argument-Based Approach: The C-test in Turkish

Researcher: Merve Demiralp (University of Bristol)

You are invited to participate in this research study. Please read the following information before you decide whether you wish to participate or not.

What is the aim of this study?
The aim of this study is to evaluate a Turkish language test as an estimate of overall language proficiency.

Why have I been invited?
You are being asked to take part in this study because you have been learning Turkish as a second language in academic settings and you have taken the Turkish Proficiency Exam conducted by Yunus Emre Institution.

What will I be asked to do if I take part in this study?
If you agree to participate, you will be asked to take a background questionnaire, a Turkish C-test, and a participant feedback survey. The Turkish C-Test consists of short Turkish reading passages in which some letters are missing in a number of words. You will try to fill in the missing letters. The study should last around 45 minutes in total.

What are the possible benefits of taking part?
If you complete the test together with questionnaires, you will receive a £10 (or $10) Starbucks gift card or 50 TL Idefix gift card through the email address you provide upon your completion of the study. Furthermore, you will practice your language skills in Turkish. Information collected in this study may benefit language researchers, teachers and testers in the future.

What are the possible risks and advantages of taking part?
There are no anticipates risks associated with this study.

What will happen if I decide not to take part or not to carry on with this study?
Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw at any time by logging out during the test. However, if you withdraw, you will not get the rewards associated with the study and your test will not be used in this study. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

Will my taking part in this study be kept confidential?
What you write during this study will remain anonymous and cannot be linked to you in any way. Study data will be protected in the researcher's personal and password protected laptop. Only the researcher will have access to the data.

What will happen to the results of the study?
The results of the study will only be used for academic purposes. The study will be a part of the researcher's dissertation study. It can also be used in academic conferences and publications.

What if there is a problem?
If you have any questions or complaints regarding this research project in general, please feel free to contact the researcher at md15381@bristol.ac.uk or her supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk. Now, if you agree to participate, please tick the consent buttons on the previous page.

Appendix 25: Instructor Information Sheet and Consent Form

Dear Participant,

My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am investigating teachers' perception of a Turkish test for foreign language learners of Turkish.

I invite you to participate in my study and take the following survey, which will take less than 30 minutes. You will receive **a £5 (or $5) Starbucks or ₺ 25 Idefix electronic gift card** if you take the survey and provide your e-mail address.

Participation in this study is anonymous and your responses will be kept completely confidential. Please read this information sheet before you participate in the study.

Thank you for your interest.

If you agree to participate, please tick the boxes below:

☐ I confirm that I have read and understood the information sheet of the study above.

☐ I agree to take part in the study above.

Participant Information Sheet

Project Title: Validating Language Tests in Second Language Acquisition Research and Educational Programs through an Argument-based Approach: The C-test in Turkish

Researcher: Merve Demiralp (University of Bristol)

You are invited to participate in this research study. Please read the following information before you decide whether you wish to participate or not.

What is the aim of this study?
The aim of this study is to evaluate a Turkish language test as an estimate of overall language proficiency.

Why have I been invited?
You are being asked to take part in this study because you have been teaching Turkish as a second language in academic settings.

What will I be asked to do if I take part in this study?
If you agree to participate, you will be asked to take a short survey your about your demographic information and views about the Turkish C-test as well as the Turkish Proficiency Exam. The survey should last less than 30 minutes.

What are the possible benefits of taking part?
If you complete the survey, you will receive a £5 (or $5) Starbucks gift card through the email address you provide upon your completion of the study. Information collected in this study may benefit language researchers, teachers and testers in the future.

What are the possible risks and advantages of taking part?
There are no anticipates risks associated with this study.

What will happen if I decide not to take part or not to carry on with this study?
Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw at any time by logging out during the survey. However, if you withdraw, you will not get the rewards associated with the study and your data will not be used in this study. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

Will my taking part in this study be kept confidential?
What you write during this study will remain anonymous and cannot be linked to you in any way. Study data will be protected in the researcher's personal and password protected laptop. Only the researcher will have access to the data.

What will happen to the results of the study?
The results of the study will only be used for academic purposes. The study will be a part of the researcher's dissertation study. It can also be used in academic conferences and publications.

What if there is a problem?
If you have any questions or complaints regarding this research project in general, please feel free to contact the researcher at md15381@bristol.ac.uk or her supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk. Now, if you agree to participate, please tick the consent buttons on the previous page.

Appendix 26: Test Taker Interviewee Information Sheet and Consent Form

Dear Participant,

My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am looking at Turkish second language learners' perception of a Turkish language test that I have developed. You are being asked to take part in this study because you are a language learner who has learned or has been learning Turkish in academic settings.

Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw the study at any time by letting me know. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

If you agree to participate, you will be asked to be interviewed about the Turkish C-test that you have taken before. The interview will be done through Skype or Zoom, and it will be recorded. The transcript of the interview will be sent to you afterwards. The interview will involve questions about your test taking and language learning experience. It will take around **30 minutes.**

What you say during the interview will remain anonymous. I will not use your names or any special information about you and you will not be identifiable in my thesis or any published material. You may also withdraw at any time by letting me know that you do not want to continue during the interview. However, if you withdraw, you will not get the rewards associated with the study. Study data will be kept in my personal and password protected laptop. Access to the study data will be protected. Only I will have access to the data.

There are no risks associated with this study. If you participate in the interview, you will also enter a **£50 Prize Draw** (Amazon gift card). Information collected in this study may benefit language researchers, teachers and testers in the future.

If you have any questions or complaints regarding this research project in general, please feel free to contact me at md15381@bristol.ac.uk or my supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk.

Thank you for your interest.

This project has been approved by the Graduate School of Education's Research Ethics Committee at the University of Bristol.

If you agree to participate, please tick the boxes below:

- ☐ I confirm that I have read and understood the information sheet of the study above.
- ☐ I consent that the interview will be recorded
- ☐ I agree to take part in the study above.

Appendix 27: Instructor Interviewee Information Sheet and Consent Form

Dear Participant,

My name is Merve Demiralp and I am a doctoral researcher at the University of Bristol. In my doctoral research, I am looking at teachers' perception of a Turkish language test that I have developed. You are being asked to take part in this study because you are an instructor of Turkish as a second language.

Participation in this study is entirely voluntary. You can choose not to participate at all or to withdraw the study at any time by letting me know. Regardless of your decision, there will be no effect on your relationship with the researcher or any other consequences.

If you agree to participate, you will be asked to be interviewed about the survey that you have taken before. The interview will be done through Skype or Zoom, and it will be recorded. The transcript of the interview will be sent to you afterwards. The interview will involve questions about your teaching experience with learners of Turkish and views of the Turkish C-test. It will take around **30 minutes**.

What you say during the interview will remain anonymous. I will not use your names or any special information about you and you will not be identifiable in my thesis or any published material. Study data will be kept in my personal and password protected laptop. Access to the study data will be protected. Only I will have access to the data.

There are no risks associated with this study. If you participate in the interview, you will also enter a **£50 Prize Draw** (Amazon gift card). Information collected in this study may benefit language researchers, teachers and testers in the future.

If you have any questions or complaints regarding this research project in general, please feel free to contact me at md15381@bristol.ac.uk or my supervisors, Dr. Shelley McKeown Jones at s.mckeownjones@bristol.ac.uk and Dr. George Leckie at g.leckie@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at soeethics@bristol.ac.uk.

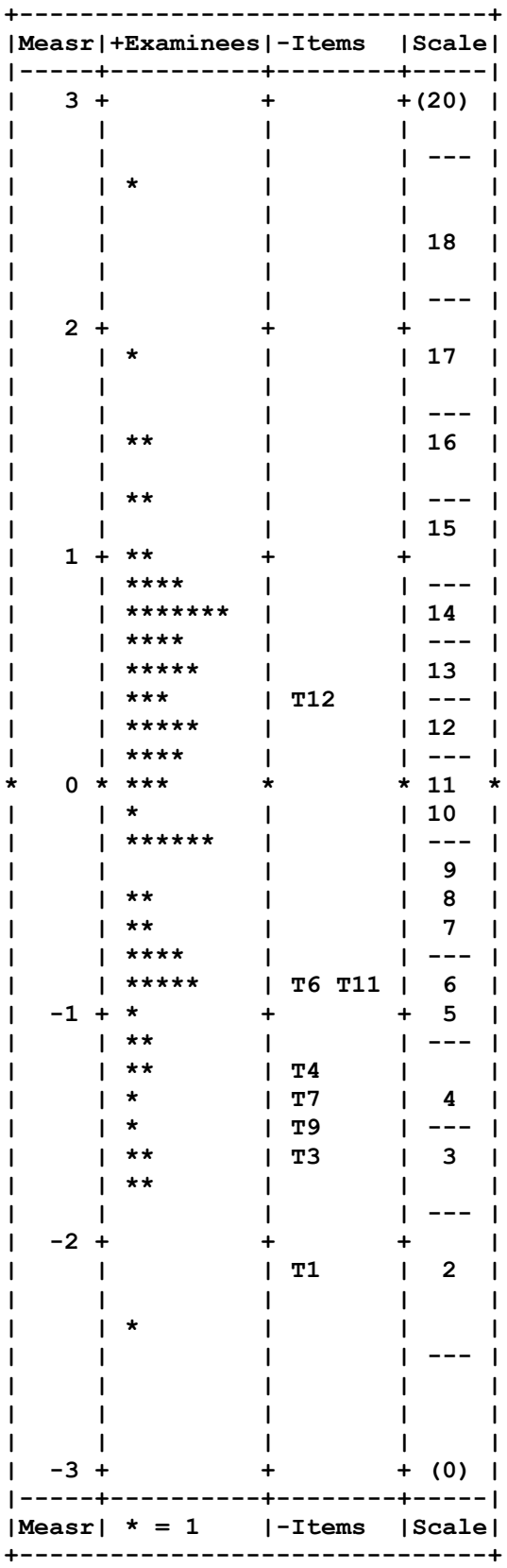
Thank you for your interest.

This project has been approved by the Graduate School of Education's Research Ethics Committee at the University of Bristol.

If you agree to participate, please tick the boxes below:

- ☐ I confirm that I have read and understood the information sheet of the study above.
- ☐ I consent that the interview will be recorded.
- ☐ I agree to take part in the study above.

Appendix 28: Rasch Analysis with 75 Examinees in Validation Study 2



Total variance explained: 84.16%
Separation=3.85, strata=5.47, reliability=.94

Text	Rpbi	Discrim	Infit	Outfit	SE	Measure
T1	.74	.88	1.07	1.12	.09	-2.05
T3	.79	1.11	.91	.88	.07	-1.55
T4	.83	1.10	.89	.86	.07	-1.25
T6	.85	1.15	.85	.91	.06	-.84
T7	.83	1.20	.85	.81	.07	-1.38
T9	.83	.70	1.50	1.34	.09	-1.45
T11	.85	1.01	.93	.95	.06	-.87
T12	.84	1.23	.78	.74	.06	.39

Appendix 29: Correlations between TYS scores and self-perceived proficiency

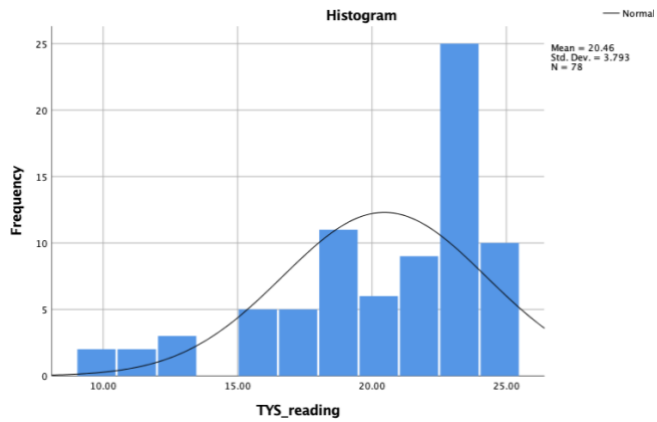
Spearman's rho between self-perceived proficiency and TYS (N=79)

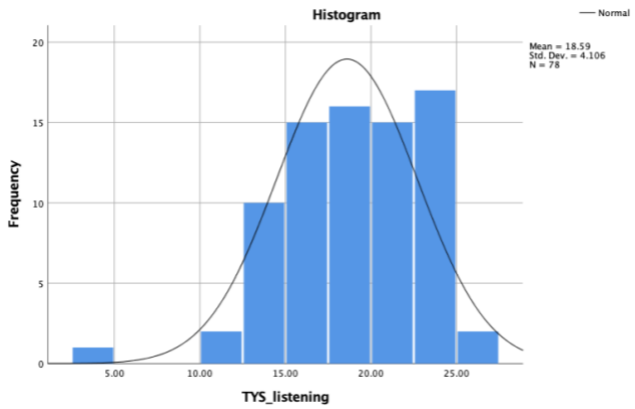
	TYS level	TYS Total	TYS reading	TYS listening	TYS speaking	TYS writing
self-reading	.60	.66	.76	.47	.42	.34
self-listening	.55	.58	.51	.58	.39	.33
self-speaking	.49	.56	.41	.37	.63	.38
self-writing	.36	.42	.25	**.18**	.32	.64
self-overall	.46	.54	.48	.38	.45	.40

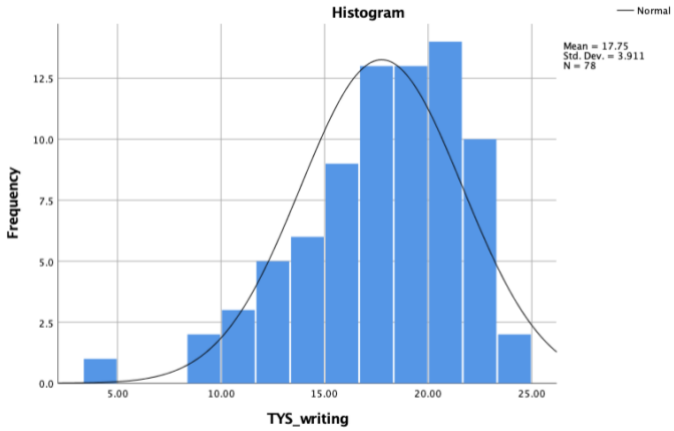
Note: all correlations (except the ones bolded) statistically significant, $p < .005$

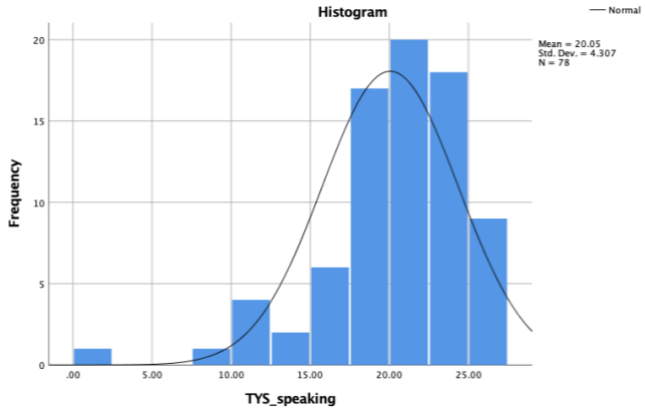
As expected, learner's self-perceived proficiency in a specific skill had the highest correlation with their score in the same skill section of TYS. For example, self-perceived proficiency in reading had the highest correlation with the TYS reading section (ρ=.76) Interestingly, self-perceived proficiency in writing had small correlations below .30 with TYS reading (ρ=.25, *p*=.03) and TYS listening (ρ=.15, *p*=10). Correlation with listening was not significant.

Appendix 30: Distribution of scores in TYS skill sections









Appendix 31: Test of Parallel Lines

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	101.167			
General	103.466[b]	.[c]	2	.

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.
b. The log-likelihood value cannot be further increased after maximum number of step-halving.
c. The log-likelihood value of the general model is smaller than that of the null model. This is because convergence cannot be attained or ascertained in estimating the general model. Therefore, the test of parallel lines cannot be performed.

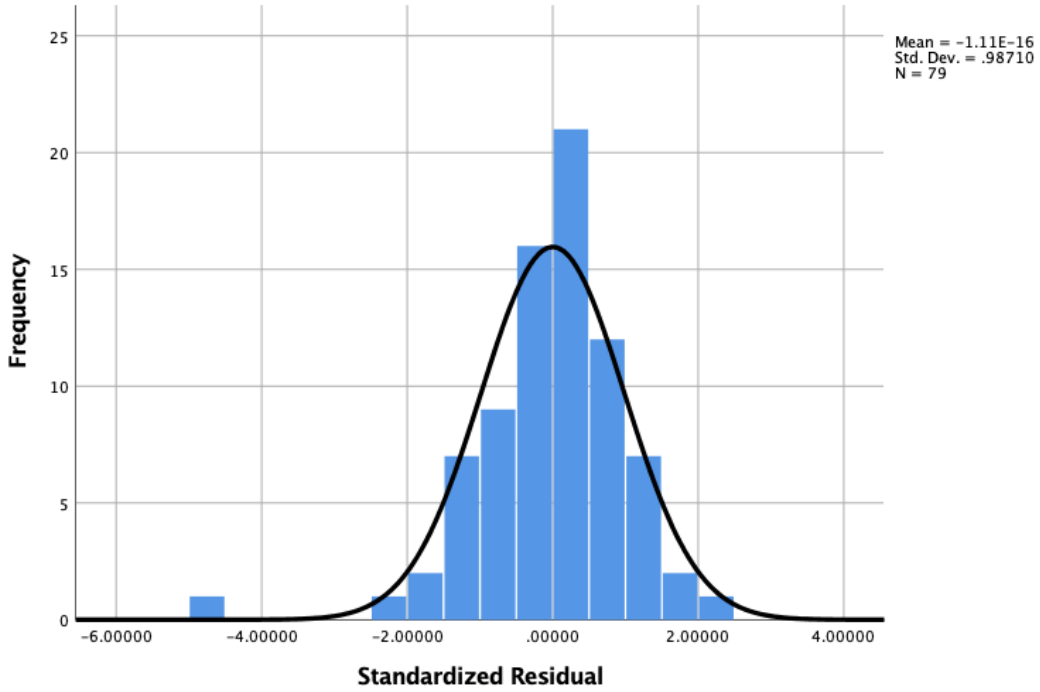
Appendix 32: Predicted TYS scores, TYS levels and C-test scores

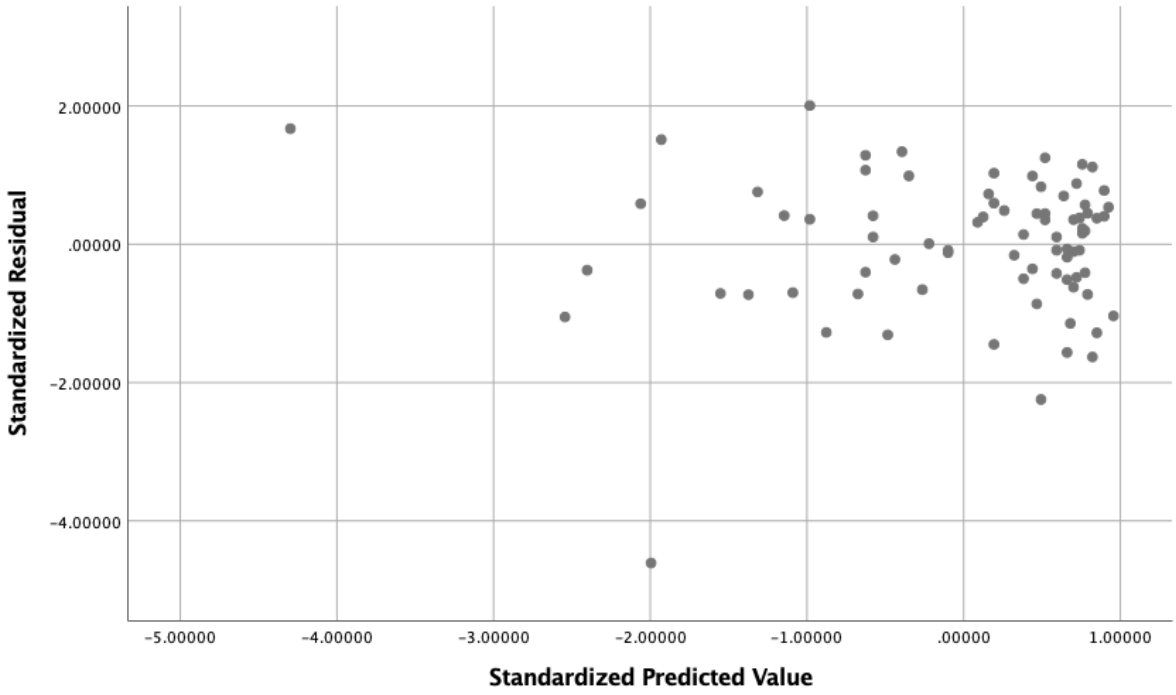
ID	Observed TYS results		Predicted TYS (linear)		Predicted TYS (quadratic)		Predicted TYS (cubic)		Predicted TYS (ordinal)	Observed C-test scores
	score	level	score	level	score	level	score	level	level	
1	82.00	C1	79.18	C1	80.77	C1	81.89	C1	C1	120
2	82.08	C1	82.31	C1	82.83	C1	83.82	C1	C1	128
3	83.75	C1	82.31	C1	82.83	C1	83.82	C1	C1	128
4	87.92	C1	76.83	C1	78.93	C1	79.82	C1	C1	114
5	82.95	C1	84.26	C1	83.88	C1	84.47	C1	C1	133
6	94.54	C2	85.43	C1	84.43	C1	84.62	C1	C1	136
7	83.50	C1	85.04	C1	84.26	C1	84.59	C1	C1	135
8	93.05	C2	81.13	C1	82.11	C1	83.22	C1	C1	125
9	69.79	B2	83.48	C1	83.48	C1	84.26	C1	C1	131
10	82.28	C1	71.36	C1	73.63	C1	73.39	C1	C1	100
11	78.41	C1	86.22	C1	84.76	C1	84.63	C1	C1	138
12	70.83	C1	63.93	B2	64.20	B2	62.58	B2	B2	81
13	85.82	C1	85.43	C1	84.43	C1	84.62	C1	C1	136
14	91.75	C2	84.65	C1	84.07	C1	84.54	C1	C1	134
15	84.95	C1	76.44	C1	78.59	C1	79.43	C1	C1	113
16	75.00	C1	69.40	B2	71.40	C1	70.70	B2	C1	95
17	78.45	C1	84.26	C1	83.88	C1	84.47	C1	C1	133
18	62.05	B2	59.24	B2	56.92	B2	55.54	B2	B2	69
19	86.29	C1	85.83	C1	84.60	C1	84.64	C1	C1	137
20	72.31	C1	69.40	B2	71.40	C1	70.70	B2	C1	95
21	84.91	C1	70.97	B2	73.19	C1	72.87	C1	C1	99
22	74.13	C1	87.78	C1	85.33	C1	84.38	C1	C1	142
23	75.00	C1	73.70	C1	76.07	C1	76.38	C1	C1	106

24	60.85	B2	70.18	B2	72.31	C1	71.79	C1	C1	97
25	74.96	C1	72.53	C1	74.88	C1	74.92	C1	C1	103
26	85.21	C1	81.13	C1	82.11	C1	83.22	C1	C1	125
27	78.22	C1	79.96	C1	81.33	C1	82.47	C1	C1	122
28	69.49	B2	65.10	B2	65.86	B2	64.36	B2	B2	84
29	83.83	C1	77.61	C1	79.57	C1	80.57	C1	C1	116
30	76.41	C1	79.18	C1	80.77	C1	81.89	C1	C1	120
31	85.00	C1	66.27	B2	67.46	B2	66.12	B2	B2	87
32	85.99	C1	81.13	C1	82.11	C1	83.22	C1	C1	125
33	74.05	C1	80.35	C1	81.59	C1	82.74	C1	C1	123
34	85.46	C1	80.35	C1	81.59	C1	82.74	C1	C1	123
35	84.13	C1	76.83	C1	78.93	C1	79.82	C1	C1	114
36	90.75	C2	90.52	C2	86.06	C1	83.05	C1	C2	149
37	81.71	C1	76.05	C1	78.25	C1	79.03	C1	C1	112
38	60.28	B2	65.49	B2	66.40	B2	64.95	B2	B2	85
39	70.62	C1	66.27	B2	67.46	B2	66.12	B2	B2	87
40	89.34	C2	89.34	C2	85.79	C1	83.76	C1	C1	146
41	86.43	C1	85.43	C1	84.43	C1	84.62	C1	C1	136
42	82.88	C1	83.48	C1	83.48	C1	84.26	C1	C1	131
43	57.26	B2	63.54	B2	63.63	B2	61.98	B2	B2	80
44	89.38	C2	83.09	C1	83.27	C1	84.13	C1	C1	130
45	75.30	C1	73.70	C1	76.07	C1	76.38	C1	C1	106
46	89.96	C2	79.96	C1	81.33	C1	82.47	C1	C1	122
47	70.80	B2	87.00	C1	85.06	C1	84.55	C1	C1	140
48	92.59	C2	89.34	C2	85.79	C1	83.76	C1	C1	146
49	73.67	C1	83.87	C1	83.69	C1	84.38	C1	C1	132
50	79.88	C1	84.65	C1	84.07	C1	84.54	C1	C1	134
51	87.58	C1	85.04	C1	84.26	C1	84.59	C1	C1	135
52	94.83	C2	87.00	C1	85.06	C1	84.55	C1	C1	140

53	89.13	C2	80.74	C1	81.86	C1	82.99	C1	C1	124
54	79.13	C1	82.31	C1	82.83	C1	83.82	C1	C1	128
55	70.83	C1	70.58	B2	72.76	C1	72.33	C1	C1	98
56	88.67	C1	86.22	C1	84.76	C1	84.63	C1	C1	138
57	81.84	C1	83.48	C1	83.48	C1	84.26	C1	C1	131
58	77.30	C1	92.47	C2	86.37	C1	81.35	C1	C2	154
59	66.25	B2	76.83	C1	78.93	C1	79.82	C1	C1	114
60	89.58	C2	85.83	C1	84.60	C1	84.64	C1	C1	137
61	62.21	B2	80.74	C1	81.86	C1	82.99	C1	C1	124
62	88.63	C1	87.78	C1	85.33	C1	84.38	C1	C1	142
63	80.68	C1	75.66	C1	77.91	C1	78.61	C1	C1	111
64	78.79	C1	78.40	C1	80.18	C1	81.26	C1	C1	118
65	80.30	C1	69.01	B2	70.93	B2	70.14	B2	C1	94
66	79.00	C1	83.48	C1	83.48	C1	84.26	C1	C1	131
67	82.19	C1	69.01	B2	70.93	B2	70.14	B2	C1	94
68	57.33	B2	67.06	B2	68.48	B2	67.29	B2	B2	89
69	81.00	C1	85.83	C1	84.60	C1	84.64	C1	C1	137
70	67.39	B2	69.01	B2	70.93	B2	70.14	B2	C1	94
71	55.67	Below B2	62.36	B2	61.89	B2	60.21	B2	B2	77
72	50.31	Below B2	57.28	B2	53.59	Below B2	52.75	Below B2	Below B2	64
73	49.73	Below B2	47.90	BelowB2	35.10	Below B2	42.23	Below B2	Below B2	40
74	68.73	B2	72.14	C1	74.47	C1	74.42	C1	C1	102
75	64.17	B2	68.62	B2	70.45	B2	69.58	B2	C1	93
76	71.45	C1	60.02	B2	58.21	B2	56.69	B2	B2	71
77	43.00	Below B2	56.50	B2	52.21	Below B2	51.67	Below B2	Below B2	62
78	17.24	Below B2	59.63	B2	57.57	B2	56.11	B2	B2	70
79	87.00	C1	84.26	C1	83.88	C1	84.47	C1	C1	133

Appendix 33: Standardized Residual Histogram and Scatterplot





Appendix 34: Classification tables for observed and predicted TYS levels

Linear Model

Observed TYS levels		Predicted TYS levels				Total
		Below B2	B2	C1	C2	
Cut scores						
Below B2	count	**1**	4	0	0	5
	%	**20%**	80%	0%	0%	100%
B2	count	0	**8**	5	0	13
	%	0%	**61.5%**	38.5%	0%	100%
C1	count	0	10	**39**	1	50
	%	0%	20%	**78%**	2%	100%
C2	count	0	0	8	**3**	11
	%	0	0	72.7%	**27.3%**	100%
Total	count	1	22	52	4	79
	%	1.3%	27.8%	65.8%	5.1%	100%

Cubic model

Observed TYS levels		Predicted TYS levels			Total
		Below B2	B2	C1	
Cut scores					
Below B2	count	**3**	2	0	5
	%	**60%**	40%	0%	100%
B2	count	0	**7**	6	13
	%	0%	**53.8%**	46.2%	100%
C1	count	0	8	**42**	50
	%	0%	16%	**84%**	100%
C2	count	0	0	11	11
	%	0	0	100%	100%
Total	count	3	17	59	79
	%	3.8%	21.5%	74.7%	100%

Ordinal Logic Model

Observed TYS levels		Predicted TYS levels				Total
		Below B2	B2	C1	C2	
Cut scores						
Below B2	count	**3**	2	0	0	5
	%	**60%**	40%	0%	0%	100%
B2	count	0	**5**	8	0	13
	%	0%	**38.5%**	61.5%	0%	100%
C1	count	0	4	**45**	1	50
	%	0%	8%	**90%**	2%	100%
C2	count	0	0	10	**1**	11
	%	0	0	90.9%	**9.1%**	100%
Total	count	3	11	63	2	79
	%	3.8%	13.9%	79.7%	2.5%	100%

Appendix 35: Classification table for C-test scores and TYS levels

C-test score	TYS_level				
	Below B2	B2	C1	C2	Total
40	1	0	0	0	1
62	1	0	0	0	1
64	1	0	0	0	1
69	0	1	0	0	1
70	1	0	0	0	1
71	0	0	1	0	1
77	1	0	0	0	1
80	0	1	0	0	1
81	0	0	1	0	1
84	0	1	0	0	1
85	0	1	0	0	1
87	0	0	2	0	2
89	0	1	0	0	1
93	0	1	0	0	1
94	0	1	2	0	3
95	0	0	2	0	2
97	0	1	0	0	1
98	0	0	1	0	1
99	0	0	1	0	1
100	0	0	1	0	1
102	0	1	0	0	1
103	0	0	1	0	1
106	0	0	2	0	2
111	0	0	1	0	1
112	0	0	1	0	1
113	0	0	1	0	1
114	0	1	2	0	3
116	0	0	1	0	1
118	0	0	1	0	1
120	0	0	2	0	2
122	0	0	1	1	2
123	0	0	2	0	2
124	0	1	0	1	2
125	0	0	2	1	3
128	0	0	3	0	3
130	0	0	0	1	1
131	0	1	3	0	4
132	0	0	1	0	1

133	0	0	3	0	3
134	0	0	1	1	2
135	0	0	2	0	2
136	0	0	2	1	3
137	0	0	2	1	3
138	0	0	2	0	2
140	0	1	0	1	2
142	0	0	2	0	2
146	0	0	0	2	2
149	0	0	0	1	1
154	0	0	1	0	1
Total	5	13	50	11	79

Note: red shows predicted below B2 level, green shows predicted B2 level, orange shows predicted C1 level.